Estimation of Population Proportion with Ranked Set Samples in the Presence of Multiple Concomitants

by

© Amirhossein Alvandi

A thesis submitted to the School of Graduate Studies in partial fulfillment of the requirements for the degree of Master of Science.

Department of Mathematics and Statistics

Memorial University

August 2019

St. John's, Newfoundland and Labrador, Canada

Abstract

Ranked Set Sampling (RSS) design as one of the most renowned cost-effective sampling methods, have found a wide range of applications from agriculture to medical research. This statistical approach is commonly used in situations where measuring a variable of interest is expensive and time-consuming, however, small ordered sets of sampling units can be produced utilizing some source of auxiliary information as the ranking criterion. These concomitant variables may include any easily measurable characteristic of the units such as visually assessed cytological characteristics of patients. When ranking the units in a set, it is often unfeasible to assign unique ranks to all units. To provide some flexibility, we allow declaring ties among the individuals, and utilize the tie structure data for estimation purposes. Several authors have applied this sampling method to the problem of population proportion estimation. Throughout this thesis, we discuss various estimation procedures for population proportion using Partially Rank Ordered Set (PROS) sampling method which accomplishes the tie declarations by dividing sampling units into partially ranked subsets. We also discuss an estimation procedure that use logistic regression proportion estimates for ranking the sampling units. However, the most important contribution of our research is to extend six non-parametric and maximum likelihood population proportion estimators to incorporate the ranking information obtained from multiple sources. Through extensive simulation studies, and real data analysis, we investigate the performance of eight proportion estimators under various settings of ranking quality, and tie structures among the ranks. We show the universal superiority of RSS-based estimators over their counterparts using Simple Random Sampling (SRS), and the significant improvement in the estimation efficiency by combining the tie structure information of multiple concomitant variables.

I dedicate this work to my family.

Lay summary

Ranked Set Sampling (RSS) is a common sampling method in situations where a characteristic of an individual is costly to measure. For example, in breast cancer studies, the determination of tumor status requires a comprehensive biopsy procedure that is expensive and time-consuming. However, a group of visually assessed cytological variables with various correlation levels with the malignancy of breast tumors is accessible. Accordingly, using these easily attainable characteristics, one can compare a group of patients regarding their probability of having malignant breast tumors. Utilizing this method, it is more likely to obtain a sample group of patients that are better representatives of the underlying population.

In ranking a group of individuals, we repeatedly deal with conditions where a characteristic of two or more patients are very close in value. In these cases we declare tie among all those patients, and select one of them randomly for the final measurement on her breast cancer status. Then the tie information of the randomly selected patients are used for increasing the accuracy of our estimation. The frequency of these cases depends on the range of values for each of the cytological variables (characteristics). However, the most important part of our research is how to benefit from all those accessible characteristics in comparing the probability of having malignant breast tumors in two or more patients. In these situations, we combine the ranking data obtained from multiple cytological variables based on their correlation level to the breast cancer status of patients.

Through numerous settings of correlation levels, group sizes, and tie frequencies, we show that the estimation accuracy uniformly increases when using RSS instead of simple random sampling. We also demonstrate the improvement in the precision of eight population proportion estimators when including the information of multiple cytological variables.

Acknowledgements

I would like to express my sincere gratitude to my supervisors, Drs. Armin Hatefi, and Taraneh Abarin for their patience, enthusiasm, motivation, and immense knowledge. I gratefully acknowledge the financial support in the form of graduate fellowships and teaching assistantships provided by Memorial University of Newfoundlands School of Graduate Studies, the Department of Mathematics and Statistics, and my supervisors.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Simple random sampling is the most common technique for data collection, however, cost-effective sampling methods are of major concern in numerous situations, especially when the measurement of the characteristic of interest is high-priced and/or time consuming. Ranked set sampling has been proved to be a successful strategy in these situations.

Ranked Set Sampling design (RSS) was first introduced by McIntyre [14] to effectively estimate the yield of pasture in Australia. However, over time and specially in recent decades, this sampling technique has been utilized for a wide range of applications including medical research, environmental and ecological studies, agriculture and forestry, to name a few. For example Wang et al. [25] used properties of ranked set samples to estimate fish stock abundance while utilizing the information of catch rate data obtained from previous years. Moreover, ranked set sampling has been employed in a broad range of fields such as medicine (Hatefi and Jafari Jozani [10], Chen et al. [2]), environmental monitoring (Chen et al. [4]), and forestry (Halls and Dell [9]).

In addition, ranked set samples have been applied to various statistical problems including the estimation of population mean (Ozturk [18]), the estimation of population variance (Perron and Sinha [19]),density function estimation(Lim et al. [12]), classification (Hatefi et al. [11]), and parametric analysis (Stokes [20]). For further information on the applications of ranked set sampling design, readers are referred to Chen et al. [24].

The basic premise for ranked set sampling scheme is the assumption that a set of sampling units drawn from the underlying population can be ranked by certain means rather cheaply and in a time efficient manner without the actual measurement of the variable of interest. By benefiting from this sampling design, we attempt to acquire a sample that is more likely to span the range of the population. Accordingly, this allows us to make legitimate statistical inferences about the questions of interest. It is worth noting that, through this thesis, the words "ranker", and "concomitant" will be used interchangeably, and with the same meaning.

Throughout this study, we examine the performance of different population proportion estimators while using ranked set sampling design. There are several factors that effects the quality of RSS technique that are taken into account in this study. Also, by incorporating the ranking quality of multiple concomitants, we may expect a decrease in the amount of ranking error. The importance of obtaining ranking information from multiple sources is more evident where we are deprived of concomitants to perform the rankings of the units confidently, and also, when we have more than one ranker with significant ranking ability. In these cases, by combining the ranking information in a meaningful manner (based on the capacity of different sources), we aim to improve the precision in estimating the population proportion.

The organization of this chapter is as follows. In Section 1.1, we primarily discuss the structure of the standard RSS, and review different characteristics of this sampling method. In Section 1.2, we discuss ties in ranking whenever assigning unique ranks to all the units in a set is not feasible. In Section 1.3, we take a critical aspect of RSS into consideration and that is incorporating the ranking potential of multiple concomitants. In Sections, 1.4, and 1.5, we provide naive estimators of the population mean and population proportion, that are simply the average of the measurements when we force the ranker to assign a unique rank to each sampling unit. Finally, in Section 1.6, we introduce a data-set (Wisconsin Breast Cancer Data) to further familiarize the reader with more sensible applications of ranked set sampling, and we summarize some of the research works that has been conducted in each of these areas.

1.1 Ranked Set Sampling Structure

In order to obtain a ranked set sample, we draw a simple random sample of size m^2 from the population, and divide these sampling units into m sets of size m. Next, before taking the final measurement, the units in each set are judgement ranked by some easily available auxiliary information. A variety of mechanisms can be used to accomplish this ranking procedure, including visual comparisons, expert opinion, or through the use of correlated concomitant variables. However, none of these methods include measuring the variable of interest. Upon ranking the sampling units in the first set, the variable of interest corresponding to the unit that is given the smallest judgment rank among the m units is measured and denoted by $Y_{[1]1}$. This is the first observation of the RSS, where square brackets are used instead of the usual parentheses for order statistics to indicate that the judgment rank and true rank may not coincide.

By using the same ranking criterion, we rank the m sampling units in the second set, and the unit with the second smallest judgement rank will be selected for taking the final measurement on its corresponding variable of interest. This becomes the second observation of the RSS, and is denoted by $Y_{[2]1}$. Then, within the third set of m units, the ranking procedure is performed in the same manner as before, and the unit with the third smallest judgment rank is identified for full measurement becoming the third observation of the RSS, say $Y_{[3]1}$. This process continues until the unit with the largest judgment rank from the m-th set is measured, and constitutes the m-th observation of the RSS, denoted by $Y_{[m]1}$. Overall, this process involves ranking of m^2 units, but taking the final measurement only on m of these units which form the first cycle of the RSS data.

Table 1.1 demonstrates the described process of obtaining one cycle of an RSS data, where $u_{sj}^{[r]}, r, s = 1, \ldots, m; j = 1, \ldots, n$ denotes the unit with the r-th (judgement) rank within the s-th set of size m, and in the j-th cycle of the standard RSS. It can be easily seen that from the s-th set of sampling units, upon performing the ranking procedure, we select the unit with the s-th smallest rank (bold-faced) for taking the final measurement. It is worth noting that we use the same ranking criterion for the ranking procedure over the entire of these sets. Then, the m observations denoted by $Y_{[r]1}, r = 1, \ldots, m;$ constitute the first cycle of the RSS data.

Table 1.1: Selection of a RSS of size m from m sets of size m (first cycle)

Set	Sampling Units		Observation
1	$\boldsymbol{u_{11}^{[1]}}, u_{21}^{[1]}, u_{31}^{[1]}, \ldots, u_{s1}^{[1]}, \ldots, u_{m1}^{[1]}$	$\Longrightarrow$	$Y_{[1]1}$
2	$u_{11}^{[2]}, \boldsymbol{u_{21}^{[2]}}, u_{31}^{[2]}, \ldots, u_{s1}^{[2]}, \ldots, u_{m1}^{[2]}$	$\Longrightarrow$	$Y_{[2]1}$
3	$u_{11}^{[3]}, u_{21}^{[3]}, \boldsymbol{u_{31}^{[3]}}, \ldots, u_{s1}^{[3]}, \ldots, u_{m1}^{[3]}$	$\Longrightarrow$	$Y_{[3]1}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
s	$u_{11}^{[s]}, u_{21}^{[s]}, u_{31}^{[s]}, \ldots, \boldsymbol{u_{s1}^{[s]}}, \ldots, u_{m1}^{[s]}$	$\Longrightarrow$	$Y_{[s]1}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
m	$u_{11}^{[m]}, u_{21}^{[m]}, u_{31}^{[m]}, \ldots, u_{s1}^{[m]}, \ldots, \boldsymbol{u_{m1}^{[m]}}$	$\Longrightarrow$	$Y_{[m]1}$

By repeating the same steps, we may construct n cycles of size m, and form a ranked set sample of the total size $N = mn$. Then, the final RSS contains mn observations denoted by

$$\left\{Y_{[r]j}, r = 1, \ldots, m; j = 1, \ldots, n\right\}.$$

We must note that each cycle contains a single copy of m (judgement) order statistics, and by replicating the same process n times, we acquire n independent copies of (judgement) order statistic from each of the m strata of the underlying population. The described scheme will result in a balanced RSS, where the descriptor balanced stands for the fact that we have collected the same number of copies of each (judgment) order statistic. It should be noted that, throughout this thesis, we focus on the balanced RSS design. On the other hand, an unbalanced RSS can be obtained by allowing the number of observations taken at each (judgment) rank to vary. This amounts to replacing n with $n_r, r = 1, \ldots, m$, where n_r denotes the number of observations taken in the r-th (judgment) stratum. Then the obtained RSS, $Y_{[r]1}, Y_{[r]2}, \ldots, Y_{[r]n_r}$, are the precise values of these n_r units, and the total sample size of this unbalanced RSS is given by

$$N = \sum_{r=1}^{m} n_r.$$

When ranking the units in each set prior to the final measurement on them, it is reasonable to expect ranking errors, and the ranking procedure is referred as imperfect

ranking. But, under perfect ranking, it is assumed that there is no ranking error involved in the construction of RSS data. As a common notation for perfect and imperfect ranking, we put the ranks in parentheses when ranking is perfect, otherwise, we put the ranks in square brackets. Thus, $Y_{(r)}$ and $Y_{[r]}$ will be generic notation for measurements with rank r when ranking is perfect and imperfect, respectively.

Set size plays an important role in the performance of any RSS design. For given set size, each measured RSS observation utilizes additional information obtained from its ranking relative to $m-1$ other units from the population. With perfect rankings this additional information is clearly an increasing function of m. Thus, under perfect ranking assumption, we would like to take our set size to be as large as economically possible within available resources. However, one might clearly see that increasing the set size will result in an increasing probability of ranking errors. This positive correlation is because in larger sets, there are greater number of candidates (units) to receive each rank r where $r = 1, \ldots, m$. Therefore, it is reasonable to investigate the impact of the set size m under imperfect ranking on the RSS statistical procedures.

1.2 Ties in Ranking

In practice, there are often cases where the ranker lacks sufficient certainty in assigning unique ranks to each of the units within a set. This issue arises based on the ranking quality when using different rankers, and also the nature of the underlying population. Now, by forcing the ranker to assign unique ranks within sets, we augment the likelihood of observing ranking errors. To tackle this issue, we allow ties in ranking process whenever the units cannot be ranked with high confidence. Then, the tie structure among the units within each set is captured for estimation purposes.

Frey [7] has studied the effect of two classes of tie structures in ranking the units within each set on the performance of various population mean estimators. Zamanzadeh and Wang [28] has investigated the behaviour of different population proportion estimators under these tie structures. Both of these tie structure classes apply the idea of perceived sizes that motivated the work of Dell and Clutter [5] on imperfect ranking in RSS. The first model, discrete perceived size (DPS), was proposed by Fligner and MacEachern [6], and the second tie structure model, tied-if-close (TIC), is relatively similar but also different in some critical ways. The two tie structure models will be

further discussed in Section 3.1.

To control the impact of ranking error, Ozturk [18] introduced partially rank ordered set (PROS) sampling designs. The PROS design controls the ranking error by increasing the set size m and reducing the number of ranked units in each set. For a given set size m, a PROS design does not force the ranker to assign unique ranks to each single unit. Instead, we allocate the units into subsets of pre-specified sizes. The units within each subset are not ranked, but each unit in subset h is considered to have smaller rank than the rank of each unit in subset h' for all $h < h'$. Arslan and Ozturk [1] developed maximum likelihood estimators for the location and scale parameters in a location-scale family of distributions. PROS sampling attempts to boost the ranker's confidence in distinguishing between the units, by partitioning each set into multiple subsets. However, this method still deals with a restriction which is the pre-specified number of subsets. Due to this assumption, the observer is forced to allocate the same number of tied units in each set. To address this issue, Frey [7] and Ozturk [16] relaxed the assumption that the number of subsets needs to be predetermined. This provides a flexibility in that the ranker is allowed to declare as many subsets as desired depending on one's ranking ability.

1.3 Combining Multiple Ranker Information

A standard RSS design benefits from the ranking information acquired from a single concomitant variable (ranker). However, in numerous studies there are access to more than one concomitant variable with different ranking abilities, and it is desirable to combine the ranking information of different concomitants to reduce any possible ranking error. This combination is a weighted average of the assigned ranks from each concomitant variable with weights determined by the ranking capacity (e.g. correlation level with the variable of interest) of different concomitant variables. Ozturk [16] introduced a mechanism to incorporate the ranking information from multiple sources in PROS setting. Ozturk [17] also studied the estimation of population mean by using multiple auxiliary variables for finite populations, and Hatefi and Jafari Jozani [10] investigated the impact of incorporating the ranking potential of multiple concomitant variables on the efficiency of PROS in the estimation of population proportion.

1.4 Estimation of Population Mean with RSS

Suppose we obtained an RSS data of size $N = mn$, $\{Y_{[r]j}, r = 1, \ldots, m; j = 1, \ldots, n\}$. The natural ranked set sample estimator, $\hat{\mu}_{RSS} = E(Y_{[r]j})$, for the population mean μ based on the balanced ranked set sample is simply the average of the sample observations, and it is given by

$$\hat{\mu}_{RSS} = \bar{Y}_{RSS} = \frac{1}{mn} \sum_{j=1}^{n} \sum_{r=1}^{m} Y_{[r]j}.$$

The balanced RSS estimator $\hat{\mu}_{RSS}$ is an unbiased estimator of the population mean regardless of whether the judgment rankings are perfect or imperfect. Dell and Clutter [5] established this result without any restriction on the accuracy of the judgment rankings, and regardless of the variability in the values of set size, and the number of cycles. Wolfe [27] discussed statistical properties of $\hat{\mu}_{RSS}$ under perfect ranking. Here, we investigate the bias, and the variance in estimation of population mean using RSS data.

Let $Y_{(1)}, \ldots, Y_{(m)}$ be the m mutually independent RSS observations under perfect ranking. It is at once apparent that the variable $Y_{(r)}, r = 1, \ldots, m$, is distributed identical to the r-th order statistic obtained from a simple random sample of size m from a continuous distribution with cumulative and density functions denoted by F, and f respectively, and with finite mean μ, and finite variance σ^2.

From properties of simple average it can be easily shown that

$$E[\hat{\mu}_{RSS}] = E[\bar{Y}_{RSS}] = \frac{1}{m} \sum_{r=1}^{m} E[Y_{(r)}], \tag{1.1}$$

Now, under the assumption of perfect ranking we have

$$E[Y_{(r)}] = \int_{-\infty}^{+\infty} y \frac{m!}{(r-1)!(m-r)!} (F(y))^{r-1} (1 - F(y))^{m-r} f(y) dy, \tag{1.2}$$

Following (1.1), and (1.2), we can obtain

$$E[\bar{Y}_{RSS}] = \frac{1}{m}\sum_{r=1}^{m}\int_{-\infty}^{+\infty}\binom{m-1}{r-1}(F(y))^{r-1}(1-F(y))^{m-r}f(y)dy \tag{1.3}$$
$$= \int_{-\infty}^{+\infty} yf(y)\sum_{r=1}^{m}\binom{m-1}{r-1}(F(y))^{r-1}(1-F(y))^{m-r}dy,$$

By replacing $r-1$ by q in (1.3), and by considering $p = F(Y)$ as the probability of success, it can be immediately observed that the summation is simply the sum over the entire sample space of the probabilities for a binomial random variable, and we may simplify the expression as follows

$$\sum_{r=1}^{m}\binom{m-1}{r-1}(F(y))^{r-1}(1-F(y))^{m-r} = \sum_{q=0}^{m-1}\binom{m-1}{q}(F(y))^{q}(1-F(y))^{(m-1)-q} = 1,$$

Then by considering (1.1), we obtain

$$E[\hat{\mu}_{RSS}] = E[\bar{Y}_{RSS}] = \int_{-\infty}^{+\infty} yf(y) = \mu. \tag{1.4}$$

Therefore, $\hat{\mu}_{RSS}$ can be considered as an unbiased estimator of the population mean. Now, to investigate the variance of $\hat{\mu}_{RSS}$, by noting the mutual independence of $Y_{(r)}, r = 1, \ldots, m$, we have

$$Var(\bar{Y}_{RSS}) = \sum_{r=1}^{m} Var(Y_{(r)}), \tag{1.5}$$

Let $\mu_{(r)} = E(Y_{(r)}), r = 1, \ldots, m$. Then we can write

$$E[(Y_{(r)}-\mu)^2] = E[(Y_{(r)}-\mu_{(r)}+\mu_{(r)}-\mu)^2] \tag{1.6}$$
$$= E[(Y_{(r)}-\mu_{(r)})^2] + (\mu_{(r)}-\mu)^2$$
$$= Var(Y_{(r)}) + (\mu_{(r)}-\mu)^2,$$

Considering that the cross product terms are equal to zero, from (1.5), and (1.6), we have

$$Var(\bar{Y}_{RSS}) = \frac{1}{m^2}\sum_{r=1}^{m} E[(Y_{(r)}-\mu)^2] - \frac{1}{m^2}\sum_{r=1}^{m}(\mu_{(r)}-\mu)^2, \tag{1.7}$$

Now, similar to what we did to prove the unbiasedness of $\hat{\mu}_{RSS}$, we can write

$$\sum_{r=1}^{m} E[(Y_{(r)}-\mu)^2] = \sum_{r=1}^{m} \int_{-\infty}^{+\infty} m(y-\mu)^2 \binom{m-1}{r-1} (F(y))^{r-1}(1-F(y))^{m-r} f(y)dy \quad (1.8)$$
$$= m \int_{-\infty}^{+\infty} m(y-\mu)^2 f(y) \sum_{r=1}^{m} \binom{m-1}{r-1} (F(y))^{r-1}(1-F(y))^{m-r} dy,$$

Again, by using the binomial distribution, the summation in (1.8) is equal to 1, and we may obtain

$$\sum_{r=1}^{m} E[(Y_{(r)}-\mu)^2] = m \int_{-\infty}^{+\infty} (y-\mu)^2 f(y)dy = m\sigma^2, \quad (1.9)$$

Then, by considering (1.7), and (1.9), we see that

$$Var(\bar{Y}_{RSS}) = \frac{1}{m^2}\left\{m\sigma^2 - \sum_{r=1}^{m}(\mu_{(r)}-\mu)^2\right\} = \frac{\sigma^2}{m} - \frac{1}{m^2}\sum_{r=1}^{m}(\mu_{(r)}-\mu)^2, \quad (1.10)$$

Therefore, $\hat{\mu}_{SRS}$, and $\hat{\mu}_{RSS}$ are unbiased estimators of the population mean. In addition, following (1.10), it can be obtained that

$$Var(\bar{Y}_{RSS}) = \frac{\sigma^2}{m} - \frac{1}{m^2}\sum_{r=1}^{m}(\mu_{(r)}-\mu)^2 = Var(\bar{Y}) - \frac{1}{m^2}\sum_{r=1}^{m}(\mu_{(r)}-\mu)^2 \leq Var(\bar{Y}). \quad (1.11)$$

Accordingly, under the assumption of perfect ranking, in addition to the unbiasedness of $\hat{\mu}_{RSS}$, this estimator has a uniformly smaller variance than its counterparts using simple random samples of the same size. We must note that the inequality in (1.11) can be considered strict unless $\mu_{(r)} = \mu$ for all $r = 1, \ldots, m$, and that is only the case when the judgement rankings are completely random.

1.5 Estimation of Population Proportion with RSS

Let Y be the variable of interest following a Bernoulli distribution with probability of success p.Suppose the estimate $\hat{p}_{RSS}$ is of our interest and the measurement of this variable is expensive and/or time consuming. Now, if we assign the numerical values of 0 and 1 to failure and success, respectively, then the proportion is simply the population average that can be obtained in the same manner as $\hat{\mu}_{RSS}$.

Terpstra [21] developed the maximum likelihood estimator $\hat{p}_{RSS,MLE}$ of population proportion. He showed that $\hat{p}_{RSS,MLE}$ is more efficient than the proportion estimator $\hat{p}_{SRS}$ obtained from a simple random sample of the same size. Terpstra and Miller [22] and Terpstra and Wang [23] examined the efficiency of several methods to construct confidence bounds for the population proportion in RSS setting. Terpstra and Liudahl[21] suggested the use of a single concomitant variable to facilitate the ranking of binary data. Furthermore, Zamanzadeh and Wang [28] have examined the performance of various non-parametric and ML estimators of the population proportion under different settings of DPS and TIC tie structures.

In population proportion estimation, logistic regression is one of the most common estimation approaches. Furthermore, when taking the final measurement on the response variable Y is rather inconvenient, Chen et al. [2] proposed to use the estimated values of probability of success under a logistic regression model fitted from a training sample. Hatefi and Jafari Jozani [10] studied the performance of logistic regression RSS and PROS proportion estimators for different combinations of concomitant variables.

The main objective of this thesis is to extend the six non-parametric, and maximum likelihood estimators of population proportion studied by Zamanzadeh and Wang [28], in a way that they would be able to include the ranking information obtained from multiple concomitants into estimation procedure.

1.6 Wisconsin Breast Cancer Data (WBCD)

Breast cancer is known as one of the most common types of cancer in women worldwide with high death rate. In practice, distinguishing between malignant and benign breast tumors requires a thorough and expensive biopsy procedure. To this end, samples of suspected masses are either surgically biopsied or verified by clinical re-examinations which takes several months following the fine needle aspiration of breast clumps. Since the early detection of cancerous breast tumors is crucial in the treatment of patients, a more efficient sampling design which reduces the frequency of the time consuming biopsy procedure is of high interest.

Wolberg and his colleagues [26], studied the Wisconsin Breast Cancer Data (WBCD), and they detected nine visually assessed cytological characteristics that are related to

Table 1.2: Cytological concomitants and their correlation with tumor status in WBCD

Concomitant	Correlation	Concomitant	Correlation
Bare nuclei	0.823	Clump thickness	0.715
Uniformity of cell shape	0.823	Marginal adhesion	0.706
Uniformity of cell size	0.821	Single epithelial cell size	0.691
Bland chromatin	0.758	Mitoses	0.423
Normal nucleoli	0.719	Subject ID	-0.058

the determination of breast cancer status in patients. These cytological concomitants can be easily obtained by examining superficial lumps or masses using the fine needle aspiration biopsy technique. This technique while being relatively inexpensive has been successfully employed for early detection of malignant breast tumors. The nine concomitant variables are ordinal variables taking on values 1 to 10, where 1 is the closest state to normal conditions and 10 represents the most abnormal case.

In Table 1.2 we display these nine concomitant variables, and the subject ID (unique code for each patient), with their Pearson correlation coefficient to the binary response variable Y, the malignancy of breast tumor. The Y values 1, and 0 are assigned to the malignant ("Success"), and benign ("Failure"), breast tumors respectively. Throughout this study, the unique ID of each patient will be used to simulate random ranking of patients.

Throughout the chapters that follows, we primarily discuss the structure of several RSS population proportion estimators, where the variable of interest Y follows a Bernoulli distribution. In Chapter 2, we describe different settings of partially rank ordered set sampling technique as well as logistic regression method for estimating population proportion. In addition, by using an illustrative example, we explain how to combine the ranking information of multiple concomitants with various ranking capacities with the variable of interest Y. In Chapter 3, by introducing three non-parametric and three ML estimators of population proportion, we extend the tie-structure of these estimators to incorporate the ranking data from all available sources. To examine the precision of these RSS estimators, in Chapter 4, by conducting a comprehensive simulation study, we investigate their performance in numerous scenarios. Finally, the methods discussed in Chapters 2, and 3 are applied to the WBCD data-set, where we estimate the prevalence of malignant breast cancer tumors in the population.

Chapter 2

Estimation Using Partially Rank Ordered Set Samples and Logistic Regression Model

When ranking units in a set, the ranker is often asked to rank all units from smallest to largest by means of concomitant variables and without actual measurement on the variable of interest. This may not be feasible in certain settings where rankers may have very little confidence to assign unique ranks to all the sampling units. However, they may be able to provide a partial ranking for the group of units in the set. For example, in a set of size four, a ranker may have high confidence to identify the smallest and largest units accurately, but may have little or no confidence to identify the second and third units. This may lead to a substantial amount of ranking error.

To tackle this issue, we study a ranking scheme in order to relax this requirement of ordering all units in each set and allow the rankers to create ordered subsets of the ranked units. The general idea is to declare some units in a set as tied whenever the ranker is unable to rank them with high confidence. These tied units are grouped into subsets. Observations for full measurement are then selected from these partially ordered subsets. Furthermore, in numerous studies, we have access to more than a single concomitant, and it is desirable to combine the ranking information from all available concomitants to reduce any possible ranking error. To this end, we attempt to investigate the partially rank ordered sampling design to incorporate the ranking information from different sources.

In Section 2.1, we introduce partially rank ordered set (PROS) sampling design. This method enables us to increase the set size without causing significant ranking error, and hence improving the information content of the sample and reducing the total cost. In ranked set sampling, we often face situations where it is difficult to rank all of the sampling units in a set with sufficient certainty, particularly when subjective information is utilized in the ranking process. Ozturk and Gao [8], and Ozturk [16] propose a judgment ordering process called judgment subsetting that allows rankers to use subsets when it is difficult to rank the entire sampling units in a set. By increasing the flexibility in ranking, they improve the precision for RSS estimation procedures under the described circumstances. Frey [7] uses partially rank-ordered sets in non-parametric population mean estimation. Ozturk [18] proposes statistical procedures that utilize partially rank-ordered data from multiple rankers to aid in selection of units for measurement in a basic ranked set sampling design. Hatefi and Jafari Jozani [10] studied the estimation of population proportion based on multiple ranker PROS sampling design. In addition, Chen et al. [2] use the logistic regression approach, and proposed an RSS-based estimation procedure for population proportion.

In Section 2.2, using the WBCD as an example, we describe how to implement the PROS sampling design with multiple concomitants to obtain a PROS sample of specified size from the population. A proportion estimator based on this sampling technique is then investigated. Finally, Section 2.3 explains a rank-based sampling method based on the estimated population proportions obtained from logistic regression model.

2.1 PROS Sampling Using a Single Ranker

When unique ranks can not be declared with high confidence, PROS sampling designs studied by Ozturk [18] can be considered as a successful approach to control ranking errors, and retain the information content of data about the population of interest. In this section, we focus on PROS sampling design. Following Ozturk [18], we initially illustrate the construction of the PROS data, and then study how population proportion can be estimated based on PROS data.

2.1.1 Construction of PROS Data Using a Single Ranker

To obtain a ranked set sample of the total size N, we draw Nm units from the population at random, and divide them randomly into N sets of m units. The individuals in each set are then ranked by the mean of some easily obtainable auxiliary variable in the same manner described in Section 1.1. It is often not realistic to expect rankers with different abilities to accurately assign unique ranks to the entire individuals within sets. But at the same time, they may be able to provide a partial ranking for the group of units in the set.

To better illustrate these situations, consider a set of $m = 5$ sampling units. Suppose a ranker may confidently identify the smallest, and the largest individuals, but can not decide about the unique ranks of the three units in the middle. In order to provide with some flexibility in ranking, as we discussed in Section 1.2, we allow the rankers to declare ties between the units. Here, the tie declaration mechanism is simply to divide individuals within each set of size m into ordered subsets. Then we may select observations for full measurement from these partially ordered subsets.

Let $S = \left\{s_{[1]}, \ldots, s_{[h]}\right\}$ be the set of these subsets that contains disjoint groups of units. Now, however the units within each subset are declared tied, the subsets are partially ranked, and that means for $l < l'$, all units in subset $s_{[l]}$ judged to have smaller ranks than all units in subset $s_{[l']}$. One might easily observe that the size of subsets are between 1 and m. This size is determined by the ranking ability of different concomitants, and also the nature of the underlying population. For example, in a set of three units $U = \{u_1, u_2, u_3\}$, the ranker may confidently identify u_3 as the largest unit, but may have little or no confidence to assign unique ranks to the other two units. In this case, we have two disjoint subsets $S = \left\{s_{[1]}, s_{[2]}\right\}$ with $s_{[1]} = \{u_1, u_2\}$ and $s_{[2]} = \{u_3\}$. Then, a unit is selected from each of these partially ordered subsets, and we only take the final measurements on these selected individuals.

To construct a Partially Rank Ordered Sets (PROS) of the total size N, we first select N sets of units, and each of size m from the population. And by applying the subsetting strategy as described to each of these N sets, we may obtain N sets of subsets

$$S_i = \left\{s_{[1]i}, \ldots, s_{[h_i]i}\right\}, i = 1, \ldots, N.$$

where $h_i; 1 < h_i < N, i = 1, \ldots, M$ denotes the number of subsets required to partition

Table 2.1: Illustration of D_1 design with $N = 5, M = 2, m = 8$

Group(i)	$S_{q,i}$	Judgement Subsets	Observation
1	$\{\boldsymbol{s_{1[1]1}}, s_{1[2]1}\}$	$\boldsymbol{\{1,2,3,4,5\}}$, $\{6,7,8\}$	$Y_{s_{1[1]1}}$
	$\{s_{2[1]1}, \boldsymbol{s_{2[2]1}}, s_{2[3]1}\}$	$\{1,2,3\}$, $\boldsymbol{\{4,5,6\}}$, $\{7,8\}$	$Y_{s_{2[2]1}}$
	$\{s_{3[1]1}, s_{3[2]1}, \boldsymbol{s_{3[3]1}}\}$	$\{1,2,3\}$, $\{4,5\}$, $\boldsymbol{\{6,7,8\}}$	$Y_{s_{3[3]1}}$
2	$\{\boldsymbol{s_{1[1]2}}, s_{1[2]2}, s_{1[3]2}\}$	$\boldsymbol{\{1,2,3\}}$, $\{4,5\}$, $\{6,7,8\}$	$Y_{s_{1[1]2}}$
	$\{s_{2[1]2}, \boldsymbol{s_{2[2]2}}\}$	$\{1,2,3,4,5,6\}$, $\boldsymbol{\{7,8\}}$	$Y_{s_{2[2]2}}$

the tied units in the set S_i. Now, We group these N sets, $S_1, \ldots, S_N$ into M different groups G_i as follows

$$G_i = \{S_{1,i}, \ldots, S_{n_i,i}\}, i = 1, \ldots, M.$$

where $\sum_{i=1}^{M} n_i = N$, and $S_{q,i} = S_j$ for some j. Moreover, $Sq, i = \{s_{q[1]i}, \ldots, s_{q[h_q]i}\}$, where $s_{q[l]i}, l = 1, \ldots, h_q$ is the l-th judgment subset of $S_{q,i}$. Then, for each group $G_i, i = 1, \ldots, M$, we select one individual for final measurement from each of the subsets $s_{l[l]i}, l = 1, \ldots, n_i$. Then the random sample from partially ranked subsets has the form

$$\{Y_{s_l[l]i}, l = 1, \ldots, n_i; i = 1, \ldots, M\}.$$

Now, we introduce three sampling designs D_1, D_2, D_3 by setting certain conditions on the construction of grouping sets. It is worth noting that in all these designs, by declaring partially ordered subsets in each set, we are able to increase the set size m for a moderate cost of ranking error.

Definition 2.1.1. *Sampling design G will be considered balanced if the grouping sets $G_i, i = 1, \ldots, M$ satisfy*

$$\bigcup_{i=1}^{M}(\bigcup_{l=1}^{n_i} s_{l[l]i}) = L\mathcal{M}.$$

where L is an integer number and $\mathcal{M}$ is the set of integers $\{1, \ldots, M\}$.

We must note that the design D_1 (as well as D_2 and D_3) is balanced since $(\bigcup_{l=1}^{3} s_{l[l]1}) \bigcup (\bigcup_{l=1}^{2} s_{l[l]2}) = 2\mathcal{M}$. Table 2.1 presents an illustrative example of constructing the PROS data under design D_1. In this example $Nm = 40$ units are identified. Then the units are placed into $N = 5$ sets, and each of size $m = 8$. In the first set of group 1, the ranker can not assign unique ranks to the entire sampling units, however, she is able to recognize that all the units in the subset $s_{1[1]1}$ have smaller ranks than the units in the subset $s_{1[2]1}$. We must also note that this partial

Table 2.2: Illustration of D_2 design with $N = 6, M = 2, m = 8$

Group(i)	$S_{q,i}$	Judgement Subsets	Observation
1	$\{\mathbf{s_{1[1]1}}, s_{1[2]1}, s_{1[3]1}, s_{1[4]1}\}$	$\mathbf{\{1,2\}}, \{3,4\}, \{5,6\}, \{7,8\}$	$Y_{s_{1[1]1}}$
	$\{s_{2[1]1}, \mathbf{s_{2[2]1}}, s_{2[3]1}, s_{2[4]1}\}$	$\{1,2\}, \mathbf{\{3,4\}}, \{5,6\}, \{7,8\}$	$Y_{s_{2[2]1}}$
	$\{s_{3[1]1}, s_{3[2]1}, \mathbf{s_{3[3]1}}, s_{3[4]1}\}$	$\{1,2\}, \{3,4\}, \mathbf{\{5,6\}}, \{7,8\}$	$Y_{s_{3[3]1}}$
	$\{s_{4[1]1}, s_{4[2]1}, s_{4[3]1}, \mathbf{s_{4[4]1}}\}$	$\{1,2\}, \{3,4\}, \{5,6\}, \mathbf{\{7,8\}}$	$Y_{s_{4[4]1}}$
2	$\{\mathbf{s_{1[1]2}}, s_{1[2]2}\}$	$\mathbf{\{1,2,3,4\}}, \{5,6,7,8\}$	$Y_{s_{1[1]2}}$
	$\{s_{2[1]2}, \mathbf{s_{2[2]2}}\}$	$\{1,2,3,4\}, \mathbf{\{5,6,7,8\}}$	$Y_{s_{2[2]2}}$

ranks have been obtained by means of an easily measurable auxiliary variable, and we have not made the final measurement on the variable of interest yet. Next, one unit will be selected at random from $s_{1[1]1}$ (bold-faced) for full measurement, and the variable of interest to this selected unit is denoted by $Y_{s_{1[1]1}}$. In the second set by declaring three subsets, the ranker selects a unit at random from the second subset $s_{2[2]1}$ (bold-faced), and then the observation will be denoted by $Y_{s_{2[2]1}}$. In a similar way, a PROS sample of size 5 is obtained, and is denoted by $\{Y_{s_{1[1]1}}, \ldots, Y_{s_{2[2]2}}\}$.

From design D_1, it is at once apparent that the ranker is able to declare any number of subsets without limitation. In other words, there is a complete flexibility that the number of subsets and subset sizes can change from one set to another set, and also from one group to another group. It should be noted that, one only needs to keep track of judgement subsets to construct a balanced grouping given in Definition 2.1.1. Now, to mitigate this challenge, we need to force some limitations in the structure of grouping sets in sampling designs D_2 and D_3.

Definition 2.1.2. *In sampling design D_2, for a given G_i, within each group set $S_{q,i}, r = 1, \ldots, n_i$, ranker is required to partition each set to subsets $s_{q[l]i}$ of the same size or equivalently $m_{li} = m_i, l = 1, \ldots, n_i; r = 1, \ldots, n_i$ where m_i denotes the number of unranked units in subset $s_{l[l]i}$.*

As shown in Table 2.2, in this sampling design, the ranker is forced to declare the same number of subsets of the same size within each group. However, the number of subsets and subset sizes can be different from one group to another group. Now, by putting further restriction on the subsetting procedure, we define sampling scheme D_3 as a special case of D_2. In the following, we primarily define D_3 sampling design, and illustrate an example of D_3 presented in Table 2.3.

Table 2.3: Illustration of D_3 design with $N = 4, M = 2, m = 8$

Group(i)	$S_{q,i}$	Judgement Subsets	Observation
1	$\{\boldsymbol{s_{1[1]1}}, s_{1[2]1}\}$	$\{\mathbf{1,2,3,4}\}, \{5,6,7,8\}$	$Y_{s_{1[1]1}}$
	$\{s_{2[1]1}, \boldsymbol{s_{2[2]1}}\}$	$\{1,2,3,4\}, \{\mathbf{5,6,7,8}\}$	$Y_{s_{2[2]1}}$
2	$\{\boldsymbol{s_{1[1]2}}, s_{1[2]2}\}$	$\{\mathbf{1,2,3,4}\}, \{5,6,7,8\}$	$Y_{s_{1[1]2}}$
	$\{s_{2[1]2}, \boldsymbol{s_{2[2]2}}\}$	$\{1,2,3,4\}, \{\mathbf{5,6,7,8}\}$	$Y_{s_{2[2]2}}$

Definition 2.1.3. *In sampling design D_3, in all sets $S_{q,i}, q = 1, \ldots, n; i = 1, \ldots, M$, the ranker is required to partition the sets into the same number of subsets in all groups $G_i, i = 1, \ldots, M$ such that each subset $s_{q[l]i}, l, q = 1, \ldots, n$ contains equal number of units(i.e. $m_{li} = m$).*

It is reasonable to expect that design D_1 would have a smaller ranking error than its counterparts D_2 and D_3 since it provides the ranker to exploit from his ability to the fullest. However, the ranking error of all these judgement subsetting processes is not significant when comparing them with a standard ranked set sample of the same size. It should be noted that, in the rest of this chapter, we focus on PROS data under design D_3, where it is assumed that each set is partitioned into the same number of subsets of equal size.

2.2 Multi-Concomitant PROS Sampling

In real life situations, there are often more than one auxiliary concomitant variables available for ranking the sampling units with different ranking capacities. For example, the WBCD data-set, introduced in Section 1.6, includes multiple concomitant variables that are substantially correlated to the malignancy of a patient's breast cancer. By focusing all of our attention on the ability of a single concomitant, we tend to ignore a significant amount of information. These auxiliary variables can be used to create judgment rank strata in an RSS design. In this section, we study the PROS sampling design where potential sampling units have a wealth of auxiliary information that can be used to rank them into partially ordered sets. Ozturk [16] developed a PROS sampling design in which the selection of units is based on the ranking information of multiple rankers. Through a meaningful combination of these

ranking information, strength-of-agreement weights are constructed based on the respective ranking ability of each ranker. These weights are then used to select a single sampling unit for precise measurement in each set. Hatefi and Jafari Jozani [10] while using multi-concomitant PROS sampling design, developed a population proportion estimator based on the PROS data. Here, we initially focus on the construction of multi-concomitant PROS sampling design. The sampling design is then presented through an illustrative example using the WBCD data-set. Finally, we discuss an estimation procedure for population proportion based on multi-concomitant PROS data.

2.2.1 Construction of PROS Data Using Multiple Rankers

In order to construct a PROS sample with multiple concomitants, we first need to specify the set size m and the number of cycles n. Let $(X,Y)^T$ denote a multivariate random variable where Y represents the variable of interest following a Bernoulli distribution with the probability of success p. Let $X = (X_1, X_2, \ldots, X_K); K \geq 1$ denote the available K concomitants. Now, we describe how to obtain a multi-concomitant PROS sample of size $N = mn$, and then we clarify the procedure by using an illustrative example from the WBCD data-set. To measure the r-th judgment order statistic, say $Y_{[r]j}(r = 1, \ldots, m; j = 1, \ldots, n)$ through this design, we draw a simple random sample of size m, $U_j^{[r]} = \left\{u_{1,j}^{[r]}, \ldots, u_{m,j}^{[r]}\right\}$ from the population. The sampling units $U_j^{[r]}, r = 1, \ldots, m; j = 1, \ldots, n$ are then ranked in each set by their respective concomitant variables $X_{k,j}^{[r]} = (X_{1,k,j}^{[r]}, \ldots, X_{m,k,j}^{[r]})$ where $X_{k,j}^{[r]}, k = 1, \ldots, K$ denotes the value of the k-th concomitant variable for each of the sampling units in $U_j^{[r]}$. Upon ranking the sampling units using $X_{k,j}^{[r]}$, the structure of rank vectors is given by

$$V_{k,j}^{[r]} = V(X_{1,k,j}^{[r]}, \ldots, X_{m,k,j}^{[r]}) = \left\{V_{1,k,j}^{[r]}, \ldots, V_{m,k,j}^{[r]}\right\}, \qquad k = 1, \ldots, K.$$

where $V_{h,k,j}^{[r]}$ is the rank assigned to the sampling unit $u_{h,j}^{[r]} \in U_j^{[r]}$. In situations where the ranker is not confident to assign unique ranks to each single unit within a set, all tied units in the set receive the same rank. Also, if there is a negative correlation between the concomitant variable and the response variable Y, the ranking procedure is accomplished by employing $V_{s,k,j}^{[r]} = m + 1 - V_{m+1-s,k,j}^{[r]}, s = 1, \ldots, m$, to produce the necessary judgment order statistics. To record the judgement ranking information

of the units of the set $U_j^{[r]}$, we construct an $m \times m$ weight matrix $W_{k,j}^{[r]}$ for each $V_{k,j}^{[r]}, k = 1, \ldots, K; j = 1, \ldots, n$. Thus the entries of the weight matrix $W_{k,j}^{[r]}$ represents the strength of weights of the ranking procedure.

It should also be noted that when no ties are declared for the h-th unit $u_{h,j}^{[r]}$, the h-th ($h = 1, \ldots, m$) row and the $V_{h,k,j}^{[r]}$-th column of the matrix $W_{k,j}^{[r]}$ is one, and the rest of entries of the h-th row is zero. If there are t tied ranks for the h-th unit ($u_{h,j}^{[r]}$), then all the entries corresponding to the tied ranks in the h-th row will be $\frac{1}{t}$ and other entries of the row will be zero. In a similar manner, we build $W_{k,j}^{[r]}$ for all $k = 1, \ldots, K$ to incorporate the ranking information obtained from all available concomitants in the selection of $Y_{[r]j}$. Then, in order to prioritize the judgement ranks obtained from concomitants with higher ranking abilities, we focus on the weighted average of the strength-of-weight matrices

$$\bar{W}_j^{[r]} = \sum_{k=1}^{K} \alpha_k W_{k,j}^{[r]}, \qquad \sum_{k=1}^{K} \alpha_k = 1. \tag{2.1}$$

The coefficient α_k represents the degree of importance of the concomitant variable, X_k, in the ranking process. It is reasonable to expect a higher precision level for the RSS estimator, when the correlation coefficient between the concomitant variable and the variable of interest, Y, is large. This correlation level between the two variables is considered as the ranking ability of different rankers. Then weight coefficients $\alpha_k, k = 1, \ldots, K$ can be computed as follows

$$\alpha_k = \frac{|\rho_k|}{\sum_{k=1}^{K} |\rho_k|}. \tag{2.2}$$

where ρ_k is the correlation coefficient between Y and $X_k, k = 1, \ldots, K$. Ultimately, since we are looking for r-th order statistic, we focus on the r-th column of $\bar{W}_j^{[r]}$. Then the unit with the maximum weight is selected for final measurement, and the variable of interest corresponding to the selected unit is denoted by $Y_{[r]j}$. If the maximum weight among the units in the r-th column is not unique, we consider all the rows with the maximum weight. For all these rows, we calculate the dispersion around the r-th judgment order statistic. Then, we select the unit with the highest dispersion for precise measurement, and identify it as $Y_{[r]j}$. Furthermore, if there are still more than one candidate upon computing their dispersion, we choose one of the units with equal dispersion at random, and we calculate its corresponding weight vector $\bar{\omega}_j^{[r]}$.

For a given weight vector $\bar{\omega}_j^{[r]} = (\bar{\omega}_j^{[r,1]}, \ldots, \bar{\omega}_j^{[r,m]}) \in \bar{W}_j^{[r]}$, the dispersion around the r-th rank stratum is given by

$$\gamma_{r,j} = \sum_{l=1}^{m} (l-r)^2 \bar{\omega}_j^{[r,l]}, \tag{2.3}$$

It can be evidently observed that small values of $\gamma_{r,j}$ suggest higher dispersion around the r-th judgement rank. At the end the observed multi-concomitant PROS sample of size $N = mn$ is given by

$$\left\{(Y_{[r]j}, \bar{\omega}_j^{[r]}), r = 1, \ldots, m; j = 1, \ldots, n\right\}.$$

Through this design, by calculating the weight vector $\bar{\omega}_j^{[r]}$, the selection of $Y_{[r]j}$ is dependent on the combination of judgement ranking information from all the available rankers. In this study, we assume that the values of the correlation coefficients are known. However, in cases where this assumption is not fulfilled, one may use any prior information such as training samples, results of previous surveys, and similar conducted studies, to estimate the unknown quantities.

Illustration of Multi-Concomitant PROS Sampling Using WBCD

In this section, we use a simple example based on the WBCD data-set, introduced in Chapter 1, with $K = 4$ cytological characteristics to show how multi-concomitant PROS data and the weight matrices are obtained. In this example, the four cytological concomitants include bland chromatin ($k = 1$), mitoses ($k = 2$), bare nuclei ($k = 3$) and finally the subject ID ($k = 4$). Suppose we are interested in measuring the third judgment order statistic $Y_{[3]1}$ when the set size is $m = 4$ and the cycle size is $n = 1$. In other words, we shall determine the status of the breast tumour of a patient ranked as the third smallest sample unit among four patients forming a set of size $m = 4$. Table 2.4 shows a set of four patients $U_1^{[3]} = \left\{u_{11}^{[3]}, u_{21}^{[3]}, u_{31}^{[3]}, u_{41}^{[3]}\right\}$ and their concomitants selected for ranking process in this example.

To illustrate the effect of the tie structure in this example, we assume that tied ranks may be declared in the ranking process using three cytological characteristics. In addition, unique ranks are assigned to the patients using the Subject ID since the ID numbers of patients are easily distinguishable. We then assume the tie structure

Table 2.4: The cytological characteristics of set of 4 randomly selected patients from the WBCD to form the set $U_1^{[3]}$.

Concomitant Variables	Patients			
	$u_{11}^{[3]}$	$u_{21}^{[3]}$	$u_{31}^{[3]}$	$u_{41}^{[3]}$
Bare nuclei	1	1	8	10
Bland chromatin	2	3	4	5
Mitoses	1	2	1	3
Subject ID	1182404	1198641	242970	255644

formed in a way so that the values of cytological concomitants are assigned into the subsets $s_1 = \{1,2\}, s_2 = \{3,4\}, s_3 = \{5,6\}, s_4 = \{7,8\}, s_5 = \{9,10\}$. To clarify, let us assume that the values of a cytological concomitant for two patients are 3, and 4. In this situation, the uncertainty in discrimination between these two units leads us to declare tie between them. For example, as shown in Table 2.4, by using the bland chromatin values for the ranking procedure, patients $u_{21}^{[3]}, u_{31}^{[3]}$ are declared tied at ranks $\{2,3\}$, patient $u_{11}^{[3]}$ is uniquely ranked 1 and patients $u_{41}^{[3]}$ receives the largest rank. Table 2.5 provides the ranks declared and the tie structure for the patients in the set $U_1^{[3]}$. Therefore, by using bland chromatin values for ranking procedure, the weights are 1/2 for the tied units $u_{21}^{[3]}, u_{31}^{[3]}$ at ranks $\{2,3\}$, 1 for $u_{11}^{[3]}$ that receives the first rank without any competing candidate, and 1 for the unit $u_{41}^{[3]}$ that has the highest bland chromatin, and assigned to the fourth rank. Using subject ID, we observe unique ranks (i.e. no tied rank) for all the patients in the last row of Table 2.5. Once the tie structures derived, we can construct the weight matrices $W_{k,1}^{[3]}, k = 1,\ldots,4$ corresponding to each concomitant variable by using the values in the Table 2.5.

Table 2.5: The tie structure for the status of patients in the set $U_1^{[3]}$.

Concomitant Variables	Patients			
	$u_{11}^{[3]}$	$u_{21}^{[3]}$	$u_{31}^{[3]}$	$u_{41}^{[3]}$
Bare nuclei	$1/2, \{1,2\}$	$1/2, \{1,2\}$	$1, \{3\}$	$1, \{4\}$
Bland chromatin	$1, \{1\}$	$1/2, \{2,3\}$	$1/2, \{2,3\}$	$1, \{4\}$
Mitoses	$1/3, \{1,2,3\}$	$1/3, \{1,2,3\}$	$1/3, \{1,2,3\}$	$1, \{4\}$
Subject ID	$1, \{3\}$	$1, \{4\}$	$1, \{1\}$	$1, \{2\}$

As indicated earlier, the rows and columns of each weight matrix stand for the units and the assigned ranks, respectively. For example, if the information of bare nuclei is used for ranking the sampling units, the tie structure is summarized in weight matrix, $W_{3,1}^{[3]}$. From Table 2.5, it is observed that unit $u_{31}^{[3]}$ is uniquely ranked 3, so the weight at row three and column three is 1 and all other entries in the third row and the third column are zero. Furthermore, since the units u_{11}^{3}, and u_{21}^{3} are declared tied for ranks $\{1,2\}$, the first and the second row of the first and second columns of $W_{3,1}^{[3]}$ will take on the value of $1/2$, and the other entries in those rows and column will be zero. We also note that unit $u_{41}^{[3]}$ is uniquely ranked 4, so the entry of the fourth row and the fourth column of $W_{3,1}^{[3]}$ is 1, and again, the rest of the entries of the fourth row as well as the other entries of the fourth column of $W_{3,1}^{[3]}$ are zero. In a similar fashion, all the weight matrices are computed as follows

$$W_{1,1}^{[3]} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \qquad W_{2,1}^{[3]} = \begin{bmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

$$W_{3,1}^{[3]} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \qquad W_{4,1}^{[3]} = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}.$$

Now, using (2.2), we can calculate the importance weight vector $\alpha = (0.363, 0.203, 0.394, 0.041)$, based on the correlations between the concomitants and the tumour status. We then compute the average weight matrix as

$$\bar{W}_1^{[3]} = \begin{bmatrix} 0.627 & 0.264 & 0.108 & 0.000 \\ 0.264 & 0.445 & 0.249 & 0.040 \\ 0.108 & 0.249 & 0.643 & 0.000 \\ 0.000 & 0.040 & 0.000 & 0.959 \end{bmatrix}.$$

The average weight matrix is then used to identify a patient in the set $U_1^{[3]}$ for the comprehensive biopsy procedure. Since our aim here is to measure the third judgment order statistic $Y_{[3]1}$, we focus on the third column of $\bar{W}_1^{[3]}$. We observe that patient

$u_{31}^{[3]}$ is the most likely candidate to be the patient with the third judgement rank of having malignant tumours in this set. Finally, the observed data based on the ranking information of these concomitants are given by

$$(Y_{[3]1}, \bar{\omega}_1^{[3]}) = (Y_{[3]1}, \{0.108, 0.249, 0.643, 0.000\}).$$

2.2.2 Multi-Concomitant PROS Proportion Estimation

Once we derived the multi-concomitant PROS data, we are able to study the problem of population proportion estimation based on the collected data denoted by $\left\{(Y_{[r]j}, \bar{\omega}_j^{[r]}), r = 1, \ldots, m; j = 1, \ldots, n\right\}$. To estimate the population proportion p that incorporates the tie information obtained from the multi-concomitant PROS sample, let $Y_{[r]j}$ be the quantified r-th judgement order statistic, and $\bar{\omega}_j^{[r]} = (\bar{\omega}_j^{[r,1]}, \ldots, \bar{\omega}_j^{[r,m]})$ be its corresponding weight vector. Hatefi and Jafari Jozani [10] proposed an estimator of the population proportion by using the multi-concomitant PROS data as follows

$$\hat{p}_{pros} = \frac{1}{m} \sum_{l=1}^{m} \frac{\sum_{r=1}^{m} \sum_{j=1}^{n} \bar{\omega}_j^{[r,l]} Y_{[r]j}}{\frac{1}{mn} \sum_{r=1}^{m} \sum_{j=1}^{n} \bar{\omega}_j^{[r,l]}}. \tag{2.4}$$

It is worth noting that $\bar{\omega}_j^{[r,l]} Y_{[r]j}$ are actually the allocation of the value of the selected unit $Y_{[r]j}$ to the l-th judgment rank, proportional to the strength of the agreement probability that l is the true rank of $Y_{[r]j}, (l = 1, \ldots, m)$. It is at once apparent that the population proportion estimator under the one-concomitant PROS data is simply a special case of (2.4).

2.3 Logistic Regression with Rank-based Sampling

Logistic Regression is the most commonly used statistical approach to study the relationship between the predictor(s) with a binary response variable. In this section, we study the RSS based estimator using logistic regression. Here, we primarily describe how the logistic regression model can be used to construct RSS data. Furthermore we discuss an RSS-based estimation procedure for population proportion using logistic regression approach. As demonstrated in Chapter 1, it is observed that the RSS estimators of the population mean are unbiased, and uniformly more efficient than

their competitors based on simple random samples of the same size. The relative efficiency of these two competitors can be obtained through the ratio of their respective variances. We must also note that the performance of RSS estimators are dependent on the error involved in the ranking procedure. In other words, the more accurately the units are ranked in each set, the estimators based on that RSS data become more efficient. It has been shown in numerous studies that the magnitude of the increase in the precision of single concomitant RSS estimator is directly contingent upon the correlation level between that concomitant variable, and the variable of interest.

An obvious candidate for ranking the units within each set in a logistic regression model is the estimated probabilities of success obtained from the model. Therefore, using easily measured characteristics of the sampling units we can rank these individuals within each set by comparing their respective estimated probabilities of success.

2.3.1 Logistic Regression Model-based Ranking Procedure

Logistic regression is frequently used to model the relationship between a binary response variable and a set of explanatory variables. Let X denote a vector of explanatory variables. The probability of success for each sampling unit is then obtained by utilizing a logistic regression model fitted from a training sample. Consider the binary logistic regression model as follows

$$logit(p) = \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta' \boldsymbol{X}, \tag{2.5}$$

where β_0 is the intercept parameter and β is the vector of slope parameters. It is straightforward to show that the probability of success, p, can be computed by

$$p = \frac{\exp(\beta_0 + \beta' \boldsymbol{X})}{1 + \exp(\beta_0 + \beta' \boldsymbol{X})}, \tag{2.6}$$

A training sample is required to construct RSS data, and achieve an RSS-based population proportion estimator using the logistic regression model. In this regard, we should have access to the values of concomitants and response variable for all the sampling units in the training data. Thus, the trained logistic regression estimate is given by

$$\hat{p} = \frac{(\exp \hat{\beta}_0 + \hat{\beta}' \boldsymbol{X})}{1 + \exp(\hat{\beta}_0 + \hat{\beta}' \boldsymbol{X})}. \tag{2.7}$$

Once the logistic regression model was trained, we can use the model to construct ranked set samples, and ranking purposes involved in the design. Consider an RSS of total size $N = mn$ with set size m, and n cycles. To construct an RSS data with multiple concomitants through logistic regression model, a set of m sampling units are first identified. Suppose we are interested in obtaining $Y_{[r]j}$ for this set. Let $p_k, k = 1, \ldots, m$ denote the probability of success, $\boldsymbol{X_k}$ represent the vector of explanatory variables for the sampling unit k within a set of size m. The estimated probability of success for this individual by using a fitted logistic regression model is obtained as

$$\hat{p}_k = \frac{(\exp \hat{\beta}_0 + \hat{\beta}' \boldsymbol{X_k})}{1 + \exp(\hat{\beta}_0 + \hat{\beta}' \boldsymbol{X_k})}, \qquad k = 1, \ldots, m. \tag{2.8}$$

The m sampling units can then be ranked based on their estimated probabilities of success, $\hat{p}_1, \ldots, \hat{p}_m$. The individual with r-th smallest probability of success is selected for full measurement, and the Y value of the corresponding measured unit is denoted by $Y_{[r]j}$. In a similar fashion, the RSS data of size $N = mn$ of the form $\left\{Y_{[r]j}, r = 1, \ldots, m; j = 1, \ldots, n\right\}$ is obtained. According to the collected binary RSS data through the logistic regression, Chen et al. [2] proposed the RSS estimator of the probability of success p as

$$\hat{p}_{l.reg} = \frac{1}{mn} \sum_{j=1}^{n} \sum_{r=1}^{m} Y_{[r]j}.$$

Throughout this chapter, by considering the importance of incorporating the information from multiple concomitants, a PROS proportion estimator was constructed in a way that it could benefit from the ranking ability of all available sources. Furthermore, since the observations have a binary distribution with probability of success p, we provided a rank-based sampling method by utilizing the estimated proportion achieved by the logistic regression method. In Chapter 4, through simulation studies and real data analysis, we investigate the effect of set size, and ranking errors on the performance of population proportion estimators that employ PROS sampling procedures as well as estimators based on logistic regression model.

Chapter 3

Non-Parametric and ML Estimation

In the construction process of RSS data, we often deal with cases where discrimination between the order of the sampling units cannot be performed with sufficient certainty. This challenge becomes more evident in studying proportion estimation where the outcomes have a binary distribution. Aside from the nature of the population under study, this problem is mostly due to the limitations in ranking ability of different concomitants. Fery [7], and Zamanzadeh and Wang [28] proposed different sampling approaches that enable the ranker to declare ties between units where required, and then use the tie structures for the estimation process upon breaking ties at random. Moreover, the randomly selected unit carries an equal amount of information on all the rank strata for which it has been declared tied. By allocating the precise values of the produced RSS data into those rank strata, we attempt to reinforce the estimation efficiency. Also, in order to obtain ordered sets of sampling units, we often have access to multiple sources of auxiliary information (e.g. different concomitant variables in the WBCD data-set) with different ranking abilities, and it is desirable to incorporate a meaningful combination of the obtained judgement ranks into our estimation procedure.

In Section 3.1 we discuss two renowned tie structure classes that are widely used to declare tie among units in a set. Section 3.2 introduces a strategy proposed by MacEachern et al. [13] to split each tied unit among the strata corresponding to the ranks for which the randomly selected individual was tied. Furthermore, throughout

Sections 3.3, and 3.4, we propose an upgraded version of the three non-parametric, and three Maximum Likelihood (ML) population proportion estimators studied by Zamanzadeh and Wang [28] while are able to utilize the combined ranking data obtained from multiple concomitants.

3.1 Tie Structure Classes in Ranking

In this section, we describe two classes of models for ties in ranking process introduced by Frey [7] including Discrete Perceived Size (DPS), and the Tied-If-Close (TIC) models. Both of these classes declare ties among the units based on the distance between their respective values for some concomitant variable X, however, they are different in some significant aspects. In a set of size m, consider the pairs $(X_r, Y_r), r = 1, \ldots, m$ as independent draws from a bivariate distribution, where Y_r, and X_r denote the variable of interest, and some easily measurable concomitant variable corresponding to the individual U_r, respectively. Then, by ranking the units using $X_1, \ldots, X_m$, the corresponding Y values are measured for the selected units. In the model introduced by Dell and Clutter [5], it is assumed that each pair of random variables $(X_r, Y_r), r = 1, \ldots, m$ satisfy

$$X_r = Y_r + \epsilon_r.$$

where $\epsilon_1, \ldots, \epsilon_m$ denote independent random variables from normal distribution with mean zero and variance σ^2. It is clear that $\sigma^2 = 0$ results in perfect ranking, where the sampling units are ranked based on the X_r values that are the same as the values of the variable of interest Y_r, and accordingly we do not deal with any ranking errors. On the other hand, for sufficiently large values of σ^2, the values of X_r, and Y_r differ considerably, and hence, the units are practically ranked at random.

3.1.1 Discrete Perceived Size (DPS)

The Discrete Perceived Size (DPS) tie structure class is based on discretizing the values of $X_r, r = 1, \ldots, m$. To do so, the ranks of the units are computed based on the values of $[X/c]$, where $[X/c]$ denotes the greatest integer smaller than X/c, and $c > 0$ is a model parameter used to control the frequency of declaring ties between the units. Thus, the DPS model declares tied rank for units U_i, and U_j, whenever their

discrete perceived sizes are identical; that is,

$$[X_i/c] = [X_j/c], \qquad i, j = 1, \ldots, m. \tag{3.1}$$

3.1.2 Tied If Close (TIC)

In this class of tie structure, tied rank is assigned to units U_i, and U_j, whenever their corresponding X values satisfy

$$|X_i - X_j| < c, \qquad i, j = 1, \ldots, m. \tag{3.2}$$

where $c > 0$ is a model parameter. In this class of tie structure, the transitivity property plays a key role since units U_i, and U_j still have the chance to receive the same rank, even in cases where the inequality (3.2) does not hold, as long as there are other units with X values that bridge the gap. It can be easily seen that when c is small, very few ties are generated between the sampling units. On the other hand, as c approaches to infinity, the probability of dealing with a tie consisting all the sampling units within each set converges to one.

By allowing ties to provide some flexibility in the construction of RSS data, we avoid a considerable amount of ranking errors, caused by random assignments of ranks. To utilize the tie structure among the individuals to improve the estimation precision, we use a special case of the splitting strategy introduced by MacEachern et al. [13] to split the final observation values among the rank strata for which the randomly selected units were declared tied. Then, using the WBCD data-set, we illustrate this splitting strategy, when incorporating the combined ranking data from multiple concomitants.

3.2 Splitting Ties into Rank Strata

As we discussed in the previous chapters, assigning unique ranks to the entire of the sampling units in a set is often unfeasible. Besides the important role of the nature of the data-set, this issue becomes more severe in cases where the set size is considerably large. In these situations, by allowing the rankers to declare as many ties as they need, the ranking errors caused by insufficient certainty of different rankers can be

considerably decreased. MacEachern et al. [13] proposed a scheme to split the ties among the sampling units in a set, and allocate the value of the variable of interest, Y, corresponding to the selected unit into all the rank strata, for which the unit has been declared tied. For example, if a unit is tied for t different ranks, then we allocate the value of Y corresponding to this unit to each of those t rank strata with weight $1/t$.

Consider an RSS data with the total sample size $N = mn$, where m represents the set size, and n is the number of cycles. The sampling procedure for the RSS using the tie information is the same as that of RSS described in Chapter 1. In here, however, the tie structures among the sampling units are recorded to increase the estimation precision. Let $U_{s[r]j}, s, r = 1, \ldots, m; j = 1, \ldots, n$ be the r-th judgement ranked unit belonging to the s-th set in the j-th cycle. Now, among the units in the r-th set, if there are more than one individual with judgement rank r, the ranker selects one of these units at random for final measurement, and records the tie structure of the r-th set in the r-th row of matrix $T^j, j = 1, \ldots, n$ that contains the tie structure of the entire sets belonging to the j-th cycle. Then the $m \times n$ tie information matrix T^j is given by

$$T^j = \begin{bmatrix} I^j_{1,1} & \cdots & I^j_{1,m} \\ \vdots & \ddots & \vdots \\ I^j_{m,1} & \cdots & I^j_{m,m} \end{bmatrix}.$$

where $I^j_{r,s}, (r = 1, \ldots, m; s = 1, \ldots, m; j = 1, \ldots, n)$ is an indicator variable defined as

$$I^j_{r,s} = \begin{cases} 1 & \text{if the unit with rank } r \text{ is tied for rank } s \text{ in the } j\text{th cycle,} \\ 0 & \text{otherwise.} \end{cases}$$

Utilizing this tie splitting strategy, Zamanzadeh and Wang [28] proposed different non-parametric, and ML estimators for the population proportion. However, the proposed estimators are only capable of employing the ranking information of a single concomitant. Yet, as we observed in the WBCD data-set, there are often different ranking criterion available. Accordingly, it is favorable to allow our estimation procedure to incorporate the ranking information, as well as the tie structures from those sources. To do so, we combine the ranking data from multiple concomitants by developing a weighted average of the tie information matrices $T^j, j = 1, \ldots, n$ corresponding to each ranker, and using their respective ranking abilities as the coefficients.

Table 3.1: The cytological characteristics of set of 5 randomly selected patients from the WBCD to form the set $U_1^{[1]}$.

Concomitant Variables	Patients				
	$u_{11}^{[1]}$	$u_{21}^{[1]}$	$u_{31}^{[1]}$	$u_{41}^{[1]}$	$u_{51}^{[1]}$
Bare nuclei	1	10	2	4	1
Clump thickness	5	5	4	6	4
Single epithelial cell size	2	7	2	3	2

Let $(X,Y)^\top$ denote a multivariate random variable, where Y represents the variable of interest following a Bernoulli distribution with probability of success p, and $X = (X_1, X_2, \ldots, X_K); K \geq 1$ denote the available K concomitant variables. By using each concomitant variable $X_k, k = 1, \ldots, K$ to rank the sampling units, when looking for the r-th judgement order statistic in the j-th cycle, (denoted by $Y_{[r]j}$), we select the unit with the maximum r-th rank from the K concomitants. In other words, we select the unit, which there is a stronger agreement on its candidacy for the r-th rank stratum. We then construct the tie structure matrices $T_k^j, k = 1, \ldots, K; j = 1, \ldots, n$ as described before, for each single ranker. Then, to combine the tie information from all the K concomitants, using (2.2), we can build a weighted average tie matrix as follows

$$\bar{T}^j = \sum_{k=1}^{K} \alpha_k T_k^j. \tag{3.3}$$

where $\sum_{k=1}^{K} \alpha_k = 1$. To clarify the process of obtaining the weighted average tie matrix $\bar{T}_k^j$, we present a simple illustrative example with the WBCD data-set. Suppose, we are interested in determining the breast cancer status of a patient which is ranked to have the smallest chance for having malignant breast tumors among a set of five patients denoted by $U_1^{[1]} = \left\{u_{11}^{[1]}, u_{21}^{[1]}, u_{31}^{[1]}, u_{41}^{[1]}, u_{51}^{[1]}\right\}$. Here, we rank the units by $K = 3$ cytological variables including bare nuclei $(k = 1)$, clump thickness $(k = 2)$, and finally, single epithelial cell size $(k = 3)$.

The values of these three cytological variables for a set of $m = 5$ patients are shown in Table 3.1. Now, by ranking the units in this set using these concomitants, we can construct the first row (which records the tie structures for $Y_{[1]1}$) of matrices $T_k^1, k = 1, 2, 3$. When using the values of bare nuclei as shown in Table3.1, the units $u_{11}^{[1]}$, and $u_{51}^{[1]}$ have the smallest value equal to 1, and are therefore declared to be tied for ranks $\{1, 2\}$. Consequently, weight 1 is assigned to the first two columns of the first

row of T_1^1, and weight 0 is assigned to the rest of the entries in this row. Individuals $u_{31}^{[1]}$, $u_{41}^{[1]}$, and $u_{21}^{[1]}$ uniquely receive the third, fourth, and fifth rank, respectively. Also, when ordering the units in the set based on the values of their clump thickness, units $u_{31}^{[1]}$, and $u_{51}^{[1]}$ receive tied ranks $\{1,2\}$.

In addition, tied ranks $\{3,4\}$ is assigned to units $u_{11}^{[1]}$, and $u_{21}^{[1]}$, since they have equal clump thickness values, and at the end the unit $u_{41}^{[1]}$ will receive the fifth rank without any competitors. Since we are interested in the individual with rank 1 in this set, and in a similar way as of T_1^1, in the first row of T_2^1, weight 1 is assigned to the first and second column, and the rest of entries in this row are equal to 0. Finally, when using single epithelial cell size as the ranking criterion, we have three equal candidates, namely $u_{11}^{[1]}$, $u_{31}^{[1]}$, and $u_{51}^{[1]}$ for receiving the smallest rank. Therefore, these three sampling units are declared tied for ranks $\{1,2,3\}$. Also, the fourth and fifth ranks will be assigned to units $u_{41}^{[1]}$, and $u_{21}^{[1]}$ with enough certainty. Thus, the first row of T_3^1 that records the tie information obtained from the third ranker, will be $(1,1,1,0,0)$.

Now, since we are looking for the first judgement order statistic in this set, it is easy to observe that unit $u_{51}^{[1]}$ has received the first rank from all three concomitants, and accordingly this unit is selected for the final measurement. It is also worth noting that if we were utilizing the ranking information of only bare nuclei, unit $u_{51}^{[1]}$ had fifty percent chance to be selected at random for the first rank stratum, since it is tied for the first rank with unit $u_{51}^{[1]}$. Similarly, if we had only the single epithelial cell size as our ranking criterion, the chance of choosing $u_{51}^{[1]}$ among the five units was 1/3. Thus, by using the strength of agreement on the judgement ranks obtained from various rankers, we are avoiding a significant amount of error in the selection procedure of sampling units. Next, we can construct the tie structure matrices $T_k^1, k = 1,\ldots,3$ to record the ties declared for the selected unit $u_{51}^{[1]}$ by each of the three rankers. Here, we consider only the elements in the first row of the tie structure matrices that are derived by ranking data within the first set.

$$T_1^1 = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix}, \qquad T_2^1 = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 \end{bmatrix}, \qquad T_3^1 = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

Now, following (3.3), we can combine the ranking data from all the three concomitants relative to their respective ranking ability. Therefore, by using $\alpha = (0.823, 0.715, 0.691)$ obtained based on the correlations between the concomitants and the breast cancer status of patients, we compute the average weight matrix as

$$\bar{T}^1 = \begin{bmatrix} 2.229 & 2.229 & 0.691 & 0.000 & 0.000 \\ 0.000 & 2.229 & 0.823 & 0.823 & 0.000 \\ 0.000 & 0715 & 2.229 & 0.691 & 0.691 \\ 0.000 & 0.000 & 1.514 & 2.229 & 0.000 \\ 0.000 & 0.000 & 0.715 & 1.538 & 2.229 \end{bmatrix}.$$

If we consider the last two elements in the first row of the weighted average tie matrix $\bar{T}^1$, since the selected unit $u_{51}^{[1]}$ has not been assigned to the fourth and the fifth ranks by any of the concomitants, it is at once apparent that the value of $Y_{[1]1}$ will not be split into those two rank strata. Moreover, it is straightforward to see that the value of $Y_{[1]1}$ will be equally allocated to the first and the second judgement rank strata, and that is because all the three concomitants declared the selected unit as tied for ranks $\{1, 2\}$. Therefore the first ranked sample, and its associated tie structure in this example is denoted by

$$\left\{Y_{[1]1}, (\bar{I}_{1,1}^1, \bar{I}_{1,2}^1, \bar{I}_{1,3}^1, \bar{I}_{1,4}^1, \bar{I}_{1,5}^1)\right\} = \left\{Y_{[1]1}, (2.229, 2.229, 0.691, 0, 0)\right\}.$$

The RSS data $Y_{[r]j}, r = 1, \ldots, m; j = 1, \ldots, n$, and the corresponding tie structure $\bar{T}^j, j = 1, \ldots, n$ can be obtained in a similar fashion.

3.3 Non-Parametric Population Proportion Estimators

Zamanzadeh and Wang [28] proposed three non-parametric population proportion estimators that utilize the tie structure obtained from only a single concomitant variable. Here, we develop those three estimators such that they can have the benefit of the combined tie structure obtained from all available sources.

The first estimator is simply the sample proportion that ignores the tie structure between the individuals in each set. However, the next two, including a split sample

proportion, and an isotonized proportion estimator, incorporate the tie information of the selected units. The standard RSS proportion estimator that disregards the ties among the individuals is obtained by

$$\hat{p}_{m.st} = \frac{1}{m}\sum_{r=1}^{m}\hat{p}_{[r]}. \tag{3.4}$$

where it is straightforward to see that $\hat{p}_{[r]} = \sum_{j=1}^{n} Y_{[r]j}/n, r = 1,\ldots,m; j = 1,\ldots,n$ is simply the sample proportion of the units belonging to the rth rank stratum. As shown in Section 1.4, $\hat{p}_{m.st}$ is an unbiased population mean estimator, and has a uniformly smaller variance relative to its counterpart based on simple random sampling.

The second non-parametric population proportion estimator we discuss here that incorporates the combined tie structure obtained from multiple concomitants is given by

$$\hat{p}_{m.sp} = \frac{1}{m}\sum_{r=1}^{m}\hat{p}_{[r]m.sp}. \tag{3.5}$$

Here the estimate $\hat{p}_{[r]m.sp}$ is the weighted average sample for the r-th rank stratum, given by

$$\hat{p}_{[r],m.sp} = \frac{\sum_{l=1}^{m}\sum_{j=1}^{n} Y_{[l]j} \times \frac{\bar{I}^{j}_{l,r}}{\sum_{s=1}^{m}\bar{I}^{j}_{l,s}}}{\sum_{l=1}^{m}\sum_{j=1}^{n}\frac{\bar{I}^{j}_{l,r}}{\sum_{s=1}^{m}\bar{I}^{j}_{l,s}}}, \tag{3.6}$$

where $\bar{I}^{j}_{l,s}$ is the entry in the l-th row and s-th column of the tie structure matrix $\bar{T}^{j}$. Therefore, by allocating the value of the variable of interest corresponding to the selected unit into all the rank strata for which this unit has been declared as tied, we enable $\hat{p}_{m.sp}$ to enjoy all segments of information comprised by every single observation.

Due to the fact that $p_{[r]} = E(Y_{[r]j})$ is the probability of success within the actual r-th rank stratum under the perfect ranking, it is expected that the estimates $p_{[r]}, r = 1,\ldots,m$ must satisfy a non-decreasing order

$$p_{[1]} \leq p_{[2]} \leq \cdots \leq p_{[m]}. \tag{3.7}$$

However, it may happen that this constraint is violated by the $p_{[r]}, r = 1,\ldots,m$ caused by the sampling variability. Following Zamanzadeh and Wang [28], and Frey [7], in these situations, we may isotonize the judgement ranked population proportion

estimates $\hat{p}_{[r]m.sp}$ by using the total weighted sample size of the r-th rank stratum with the weights designed as follows

$$n_r = \sum_{l=1}^{m}\sum_{j=1}^{n}\frac{\bar{I}_{l,r}^{j}}{\sum_{s=1}^{m}\bar{I}_{l,s}^{j}}.$$

Now, by considering $p(r)$ as the probability of success of the statistic $Y_{(r)}, r = 1,\ldots,m$ under perfect ranking, we attempt to minimize the following weighted least squares under the constraint (3.7)

$$\sum_{r=1}^{m} n_r(p_{[r]} - \hat{p}_{[r]m.sp})^2.$$

Finally, by defining $n_{\nu\eta} = \sum_{w=\nu}^{\eta} n_w$, and following Wang et al. [25], the isotonized regression estimators of $p_{[r]}, r = 1,\ldots,m$ is given by

$$\hat{p}_{[r],m.iso} = \min_{1\le\nu\le r}\max_{r\le\eta\le m}\sum_{w=\nu}^{\eta}\frac{n_w\hat{p}_{[w]m.sp}}{n_{\nu\eta}}.$$

Then the third non-parametric population proportion estimator established by the isotonized versions of $\hat{p}_{[1]m.sp},\ldots,\hat{p}_{[m]m.sp}$ are obtained by

$$\hat{p}_{m.iso} = \frac{1}{m}\sum_{r=1}^{m}\hat{p}_{[r]m.iso}. \tag{3.8}$$

It should be noted that, although the population proportions $\hat{p}_{m.sp}$, and $\hat{p}_{m.iso}$ may seem similar to non-parametric estimators in Zamanzadeh and Wang [28] at first glance, here, the tie structure, and RSS data collection are significantly different. These new estimators are capable of using tie structures, and ranking information of multiple concomitants.

3.4 Likelihood-based Estimators

Zamanzadeh and Wang [28] proposed three ML population mean estimators based on the RSS data. Despite the fact that the proposed estimators are able to incorporate the tie information in estimation, they lack the advantage of utilizing the combined tie structure from multiple resources. To deal with this drawback, we developed their proposed ML proportion estimators, so that they can benefit from the tie structure

information of any number of concomitants. To construct these new ML estimators, we need to find the distribution of the obtained RSS data from a Bernoulli population. We initially focus on probability of success of the r-th order statistic. Let $\{Y_1, \ldots, Y_m\}$ be a random sample of size m from a Bernoulli population with the probability of success p. Due to the binary nature of the population, the r-th order statistic $Y_{(r)}$ follows a binary distribution with probability of success $p(r)$. Now, if the r-th order statistic $Y_{(r)} = 1$, it is straightforward to show that at least $m - r + 1$ units must be equal to 1. Then we have

$$\begin{aligned} p(r) &= Pr(Y_{(r)} = 1) = Pr(\text{at least } m-r+1 \text{ of } Y's \text{ are } 1) \\ &= \sum_{t=m-r+1}^{m} \binom{m}{t} Pr(Y=1)^t (1 - Pr(Y=1))^{m-t} \\ &= \sum_{t=m-r+1}^{m} \binom{m}{t} p^t (1-p)^{m-t} \equiv B_{(m-r+1,r)}(p) \end{aligned} \tag{3.9}$$

where $B_{m-r+1,r}(p)$ is the incomplete cumulative distribution function of the beta distribution with parameters $m - r + 1$ and r that is evaluated at the point p, given by

$$B_{m-r+1,r}(p) = \int_0^p (m-r+1)\binom{m}{m-r+1} t^{m-r}(1-t)^{r-1} dt,$$

Let $\left\{Y_{(r)j}, r = 1, \ldots, m; j = 1, \ldots, n\right\}$ be an RSS data under the perfect ranking (i.e by using the true values of Y's for the ranking purposes, and consequently in the absence of ranking error). Then, by denoting that $Y_{(r)j} \sim Bernoulli(p_{(r)})$ where $p_{(r)} \equiv B_{m-r+1,r}(p)$, the log-likelihood function can be expressed as

$$l(p) = \sum_{r=1}^{m}\sum_{j=1}^{n} \left\{Y_{(r)j} \log p_{(r)} + (1 - Y_{(r)j}) \log\left(1 - p_{(r)}\right)\right\}, \tag{3.10}$$

where the log-likelihood function depends on the probability of success p, through the probability of success of the r-th order statistic. Therefore, the first ML estimator based on RSS is given by

$$\hat{p}_{m.ml} = \underset{p\in[0,1]}{\operatorname{argmax}}\, l(p). \tag{3.11}$$

Note that the existence of $\hat{p}_{m.ml}$ is trivial as $L(p)$ is concave in p over the unit interval. In general however, $\hat{p}_{m.ml}$ does not exist in closed form, but it can easily be computed using numerical methods. It is at once apparent that the ML estimator in (3.11) is

not able to use the tie structure information stored in RSS data. Therefore, we need to propose another ML estimator with such quality.

Suppose $Y_{(r)j}$ are the RSS observations that has been declared tied by multiple concomitants, and for several ranks in a set of m units. Then $Y_{(r)j} \sim Bernoulli(p_{(r),t})$, where $p_{(r),t}$ denotes the probability of success of the r-th order statistic when incorporating the combined tie structure information. To find $p_{(r)t}$, we define a latent variable $\Delta_j^{[r]} = (\delta_j^{[r,1]}, \ldots, \delta_j^{[r,m]})$, where $\delta_j^{[r,l]}, r = 1, \ldots, m; l = 1, \ldots, m, j = 1, \ldots, n$ is an indicator variable defined as

$$\delta_j^{[r,l]} = \begin{cases} 1 & \text{when } Y_{(r)j} \text{ is selected from the } l\text{-th tied rank,} \\ 0 & \text{otherwise.} \end{cases}$$

Considering that $Y_{(r)j}$ is selected randomly from one of the tied ranks, it is straightforward to see that $\sum_{l=1}^{m} \delta_j^{[r,l]} = 1$, and we have

$$\begin{aligned} p_{(r),t} = Pr(Y_{(r)j} = 1) &= \sum_{\delta} p_{Y_{(r)},\Delta_j^{[r]}}(\delta_j^{[r]}) \\ &= \sum_{\delta_j^{[r,1]},\ldots,\delta_j^{[r,m]}=1} \prod_{l=1}^{m} \left\{ \frac{\bar{I}_{r,l}^{j} p_{(r)}}{\sum_{l=1}^{m} \bar{I}_{[r,s]}^{j}} \right\}^{\delta_j^{r,l}} \\ &= \frac{\sum_{l=1}^{m} \bar{I}_{r,l}^{j} p_{(r)}}{\sum_{l=1}^{m} \bar{I}_{r,l}^{j}} \end{aligned} \tag{3.12}$$

where $p_{(r)}$ is given by (3.9). Then the log-likelihood function of $Y_{(r)j}$ that incorporates the tie structure information among the sampling units is

$$l_t(p) = \sum_{r=1}^{m} \sum_{j=1}^{n} \left\{ Y_{(r)j} \log p_{(r),t} + (1 - Y_{(r)j}) \log (1 - p_{(r),t}) \right\}, \tag{3.13}$$

Then, the second ML proportion estimator based on this method can be computed by

$$\hat{p}_{t.m.ml} = \underset{p \in [0,1]}{\operatorname{argmax}}\, l_t(p). \tag{3.14}$$

In addition, in the case of perfect ranking, a pseudo-likelihood function can be presented in which $Y_{(r)j}$ can be replaced by $n_r \hat{p}_{[r],m.sp}$. Finally the third log-likelihood

function is given by

$$l_{m.sp}(p) = \sum_{r=1}^{m} \left\{ n_r \hat{p}_{(r),m.sp} \log p_{(r)} + n_r (1 - \hat{p}_{(r),m.sp}) \log (1 - p_{(r)}) \right\}, \tag{3.15}$$

From (3.15), the pseudo-ML estimates of the population proportion $\hat{p}_{m.m.ml}$ that combines the strength from the ML estimation for binary data, and the tie splitting strategy described in Section 3.2 can be obtained by

$$\hat{p}_{m.m.ml} = \underset{p \in [0,1]}{\operatorname{argmax}}\, l_{m.sp}(p). \tag{3.16}$$

Now, we introduce log-concave functions to examine the uniqueness of the ML estimates obtained by maximizing the three log-likelihood functions.

Definition 3.4.1. *A non-negative function $f(x)$ defined on an interval (a, b) is said to be logarithmic concave, if for every $x, y \in (a, b)$ and every $0 \leq \lambda \leq 1$, we have*

$$f(\lambda x + (1 - \lambda) y) \geq [f(x)]^{\lambda} [f(y)]^{(1-\lambda)}.$$

It is straightforward to show that the incomplete Beta function $B_{m-r+1,r}(p)$ is a log-concave function on its domain (p), and accordingly, the log-likelihood functions $l(p)$, and $l_{m.sp}$ will be log-concave as well. Furthermore, from Mu [15], it can be concluded that the statistic $p_{(r),t}$ given in (3.12) is strictly log-concave, and hence the log-likelihood function $l_t(p)$ is strictly log-concave. Finally, from the log-concavity of the three log-likelihood functions, it can be derived that all three ML estimates can be easily obtained by using standard optimization methods.

In addition, it should be noted that the three ML estimators are obtained under the assumption of perfect ranking in which there is no ranking error involved in the construction of RSS data. However, in real life applications, the ranking error is often inevitable. For example, in the WBCD data-set the individuals are ranked based on their corresponding values for a group of cytological characteristics. For that reason, in Chapter 4, we study the performance of these three ML proportion estimators under imperfect ranking using various ranking abilities, and investigate the robustness of these methods in the presence of ranking errors.

Throughout this chapter, we initially discussed two tie structure classes that formed the basis of our RSS data generation in Chapter 4. Then by describing the

tie splitting strategy for breaking the ties among the units at random, we established a mechanism to combine the tie structure information obtained from any number of rankers. Finally, we proposed three non-parametric and three ML estimators of population proportion that utilize the ranking information of multiple concomitants. The numerical studies and real data analysis of this chapter will be widely discussed through Section 4.3.

Chapter 4

Numerical Studies

In this chapter we examine the performance of the proposed population proportion estimators in different settings. To achieve this goal, primarily, we evaluate the efficiency of each estimator by conducting a thorough simulation study that covers various aspects of the problem. It should be noted that different estimators have been proposed in the previous chapters, and accordingly throughout this chapter, we are going to assess the efficiency of a considerably large number of estimators. Chapter 2 introduced population proportion estimators based on PROS data as well as estimators using the logistic regression model. In Chapter 3, we introduced three non-parametric, and three ML proportion estimators to enjoy the ranking information from all available sources. The estimation procedures will then be applied to the Wisconsin Breast Cancer Data (WBCD), introduced in Chapter 1, for analysis of disease prevalence through the population.

In Section 4.1, we discuss the common simulation settings for all eight proportion estimators introduced in the previous chapters. Next, in Section 4.2, we examine the performance of proportion estimator based on the PROS data, and the logistic regression-based estimator. As indicated in Chapter 2, these two estimators benefit from the ranking ability of concomitant variables with different correlation levels with the binary response variable. Throughout Section 4.3, the efficiency of six non-parametric, and ML estimators that has been discussed in Chapter 3 will be investigated extensively. Moreover, through both of these sections, we measure the improvement in estimation precision by employing the ranking information from a single ranker as well as including the ranking abilities from multiple sources. Finally,

in Section 4.4, we provide an application of RSS-based estimation procedure in real life situations using the WBCD data-set.

4.1 Simulation Setup

In order to measure the performance of the previously proposed estimators, let Y be a binary variable with the probability of success p, and X a continuous random variable such that $X \sim N(\mu_x, {\sigma_x}^2)$. Note that random variables X and Y represent the concomitant variable and the binary variable of interest respectively. To generate the values of the random variable X based on a correlation level ρ (ranking ability) with the variable of interest Y, we follow the strategy of Zamanzadeh and Wang [28]. Let the conditional distribution of X given Y follow a normal distribution, and generate the values of X from conditional distribution $X|Y$ given by

$$X|Y = y \sim N(\mu_y, 1), \tag{4.1}$$

where Y takes on the values zero, and one. Now, since the conditional variance of X given $Y = y$ is assumed to be one, we may see that

$$Var(X|Y = y) = {\sigma_X}^2(1 - \rho^2) = 1, \tag{4.2}$$

Therefore, the marginal standard deviation of X is given by

$$\sigma_X = \frac{1}{\sqrt{1 - \rho^2}}, \tag{4.3}$$

Accordingly, to generate the values of concomitant variable X from (4.1), we need to find $\mu_y = E[X|Y = y]$ for $y = 0, 1$. Let $\mu_0 = E[X|Y = 0] = 0$. Then, it is straightforward to show that

$$\mu_y = E[X|Y = y] = \mu_x + \rho\frac{\sigma_x}{\sigma_y}(y - \mu_y), \tag{4.4}$$

Combining (4.3), and (4.4), we can observe that

$$\mu_y = E[X|Y = y] = \mu_x + \frac{\rho}{\sqrt{(1 - \rho^2)p(1 - p)}}(y - E(y)),$$

By setting $\mu_0 = 0$, we obtain

$$\mu_0 = \mu_x - \frac{\rho p}{\sqrt{(1-\rho^2)p(1-p)}} = 0.$$

Hence,

$$\mu_x = \frac{\rho p}{\sqrt{(1-\rho^2)p(1-p)}}, \tag{4.5}$$

Next, from (4.3), it can be easily observed that

$$\begin{aligned}\mu_1 &= \mu_x + \frac{\rho(1-p)}{\sqrt{(1-\rho^2)p(1-p)}} \\ &= \frac{\rho p}{\sqrt{(1-\rho^2)p(1-p)}} + \frac{\rho(1-p)}{\sqrt{(1-\rho^2)p(1-p)}} \\ &= \frac{\rho}{\sqrt{(1-\rho^2)p(1-p)}}.\end{aligned} \tag{4.6}$$

Here, the correlation level between X and Y defines the ranking ability of each ranker. To draw a holistic picture of the performance of the previously proposed estimators, we examine their efficiency based on different values of the probability of success p. To this end, we let $p \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$, and from the symmetry property of the Bernoulli distribution, the same behaviour is observed from all estimators for the proportion values $p \in \{0.5, 0.6, 0.7, 0.8, 0.9\}$. We perform the rankings in each set with respect to three concomitant variables with correlations $\rho = 0.9$, $\rho = 0.7$, and $\rho = 0.5$ representing strong, fair, and poor ranking capabilities. By generating $20,000$ samples of size $N = 15$ while using (4.6), we fix the total sample size to be $N = 15$. To investigate the effect of different set sizes, in one case, the set size is $m = 3$, and the cycle size $n = 5$, and in the other case, we set $m = 5$, and $n = 3$.

Over the entire of this chapter, we compare the performance of different proportion estimators by using their relative efficiencies with respect to the population proportion estimator based on SRS data of the same size. To do so, the relative efficiency of an RSS estimator can be computed by

$$RE(\hat{p}_{RSS}) = \frac{MSE(\hat{p}_{SRS})}{MSE(\hat{p}_{RSS})}. \tag{4.7}$$

It is straightforward to observe that the $RE(\hat{p}_{RSS})$ values greater than one, imply that the RSS based estimators have higher precision in estimating population proportion

relative to their counterpart using simple random sampling method with the same sample size.

4.2 Estimation based on PROS Data and Logistic Regression

In the first simulation study, we examine the effect of different factors on the efficiency of these two estimation procedures by using different settings of ranking ability of concomitants, set size, and number of cycles.

In order to utilize the tie information from different concomitants, rankers declare tie among the units by using the DPS tie structure model with parameter $c = \delta/m$, where δ denotes the range of concomitant variable values corresponding to the units within each set of size m. Also, through assigning the same number of categories to the subsets, we acquire roughly balanced PROS samples under the assumptions of D_3 sampling design introduced in Section 2.1. We also have used training samples of size $N = 15$ for the estimator $\hat{p}_{l.reg}$ to estimate the logistic regression model parameters.

Figure 4.1 shows the relative efficiencies of population proportion estimates using PROS data ($\hat{p}_{pros}$), and logistic regression model ($\hat{p}_{l.reg}$) to their counterparts based on simple random sampling of the same size $N = 15$. The relative efficiencies associated with $\hat{p}_{pros}$ are presented in blue, and the relative efficiencies of $\hat{p}_{l.reg}$ are presented with red colour. In order to investigate the effect of ranking information on the estimation performance, we provide relative efficiencies when when a single ranker is employed to rank the units, and also when the ranks are assigned by multiple concomitants.

Here, solid line demonstrates relative efficiencies of estimators based on a single concomitant variable, with different correlation coefficients with the binary variable of interest. Dashed line indicate the relative efficiencies of estimators while incorporating the ranking data from two concomitant variable with the same correlation coefficient with the variable of interest. And finally, dotted lines is used to show the relative efficiencies of estimators using three concomitant with same correlation level with the binary response variable.

It is apparent that the relative efficiencies of both estimators are greater that one in virtually all the cases, and accordingly, both estimators perform better than $\hat{p}_{SRS}$

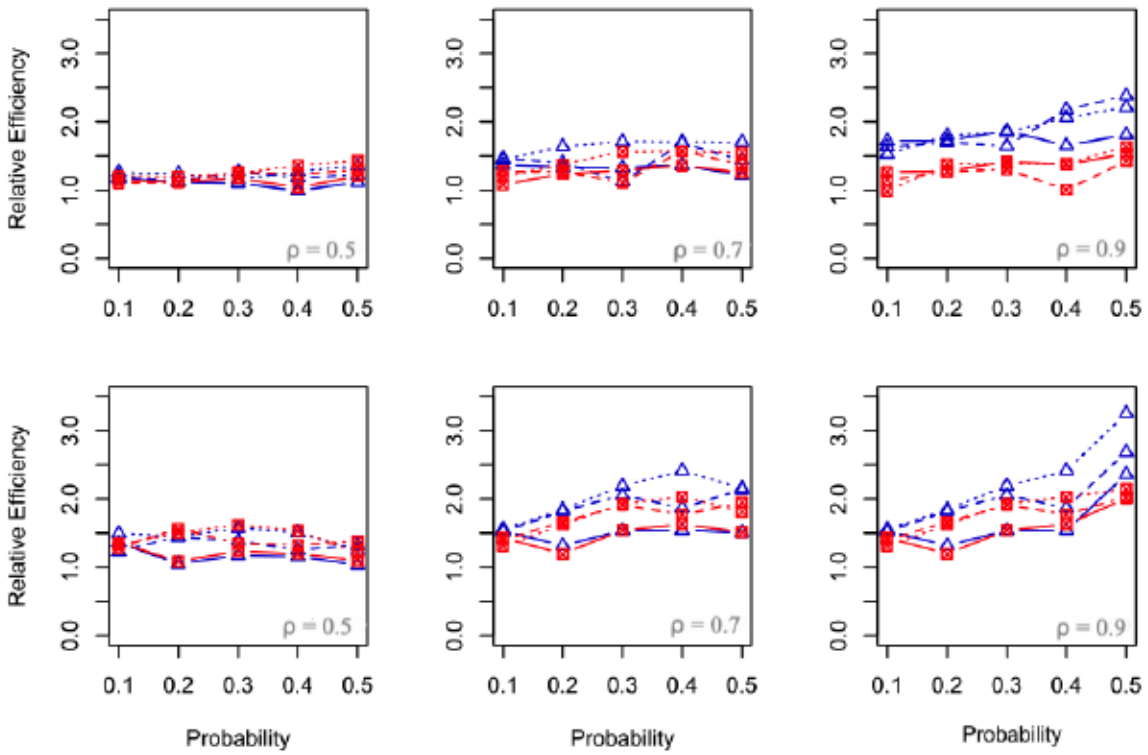

Figure 4.1: Relative efficiencies of $\hat{p}_{pros}$ (represented by △), and $\hat{p}_{l.reg}$ (represented by ⊠) using set size $m = 3$, and $n = 5$ cycles (Upper Panel), and set size $m = 5$, and $n = 3$ cycles (Lower Panel). The solid line, dashed line, and dotted line represent models with a single, two, and three concomitant(s) respectively.

even when the ranking quality is considerably low ($\rho = 0.5$). From Figure 4.1, it can be perceived that the $\hat{p}_{pros}$ outperforms $\hat{p}_{l.reg}$ in nearly all different settings with respect to their relative efficiencies. The advantage in the performance of $\hat{p}_{pros}$ is clearer for higher correlation levels. Considering the results for two different set sizes, we can observe that the relative efficiencies of both estimators have improved by increasing the set size from $m = 3$ to $m = 5$, while the total sample size is fixed at $N = 15$. The improvement in the performance of $\hat{p}_{pros}$ is due to the less amount of restrictions in the construction of subsets for obtaining a PROS data for larger values for the set size. Also, we may observe from Figure 4.1, that the effect of set size values is more conspicuous when we employ multiple concomitants into the estimation procedure.

By focusing on the obtained relative efficiencies for the two estimators when utilizing the ranking abilities of different number of concomitants, it can be clearly noticed that, in virtually all the cases, both proportion estimators have achieved higher relative efficiencies while employing multiple concomitants. We can also see that by

increasing the number of rankers, however the precision of $\hat{p}_{reg}$ improves considerably at some certain values of p, the overall increase in the relative efficiency of $\hat{p}_{pros}$ is more noticeable. Finally, we can see from Figure 4.1 that the relative efficiency patterns for $\hat{p}_{l.reg}$ is not stable for some proportion values p, and this may be rooted in the effect of training samples with measurements on both the response variable and the concomitant variables that are required for estimating the logistic regression model parameters.

4.3 Non-parametric and ML Estimation

In Chapter 3, we introduced various non-parametric and ML estimators of population proportion. The non-parametric estimators include the standard RSS-based estimator $\hat{p}_{m.st}$, the RSS estimator using the splitting strategy to break ties at random $\hat{p}_{m.sp}$, and its isotonized version $\hat{p}_{m.iso}$. The maximum likelihood estimators proposed for population proportion estimation include standard ML estimator that ignores the tie among units $\hat{p}_{m.ml}$, the ML estimator based on tie structure $\hat{p}_{t.m.ml}$, and its modified version $\hat{p}_{m.m.ml}$.

To further illustrate the effect of combining the ranking information from multiple concomitants on the efficiency of the proportion estimators in Chapter 3, we are required to broaden the scope of our simulation study. This includes investigating the performance of these estimators based on the number of concomitants employed in the ranking procedure, different ranking abilities, the effect of tie structures, and the role of different values of set size.

In order to simulate different mechanisms to declare ties between units in each set, we utilize the Discrete Perceived Size (DPS) and Tied If Close (TIC) tie structure classes described in Section 3.1. In addition, we examine the tie declaration quality while comparing the results for both the DPS, and the TIC classes at certain values of the model parameter denoted by c. At each setting of factors effecting the estimation efficiency, we discuss the impact of incorporating the ranking quality of multiple concomitants on the performance of the proportion estimators. Extensive simulation studies are carried out to fully evaluate the precision of these six estimators.

To this end, we use the same method of data generation introduced in Section 4.1, however, here to simulate the tie structure inside each set, we generate RSS

data using the DPS, and the TIC models with parameter $c \in \{0.5, 1, 4\}$ which can be considered the separation level among the units. We also study the impact of adding more concomitants into the ranking procedure, by using different correlation levels between concomitant variables, and the binary variable of interest $\rho \in \{0.5, 0.7, 0.9\}$. We also studied the effect of adding more rankers with similar ranking capacity to the model. In all the Figures 4.2 to 4.13, the left panel of each figure indicates the relative efficiency of the six estimators $\{\hat{p}_{m.st}, \hat{p}_{m.sp}, \hat{p}_{m.iso}, \hat{p}_{m.ml}, \hat{p}_{t.m.ml}, \hat{p}_{m.m.ml}\}$ utilizing the ranking information of a single concomitant variable with different ranking abilities. The middle panel and the right panel represent the efficiency of the six estimators by incorporating the ranking information from two, and three concomitants with same correlation level with the variable of interest respectively.

Primarily, it is apparent that over the entire simulation settings, the six RSS-based estimators have reached relative efficiencies greater than one. Thus, similar to what we observed for $\hat{p}_{pros}$, and $\hat{p}_{l.reg}$, all these non-parametric, and ML proportion estimators have stronger precision level relative to those based on simple random sampling.

By focusing on the results for the two different set size values, it can be clearly observed that the relative efficiencies are considerably greater when we divide the units in sets of size $m = 5$, especially under the DPS tie structure model. This increase in the estimation efficiency is the consequence of a larger number of rank strata for larger sets, and the samples are more likely to span a wider range of observations from the underlying population, although, for significantly large values of set size m, we may expect larger amount of ranking errors, and accordingly a poor estimation precision.

We can also see that the two ML estimators that incorporate the tie structure among the units $\hat{p}_{t.m.ml}$, and $\hat{p}_{m.m.ml}$ have better relative efficiencies than the ML estimator that ignores such information $\hat{p}_{m.ml}$ throughout nearly all the simulation settings. Also, the relative efficiency of $\hat{p}_{t.m.ml}$ is more stable than $\hat{p}_{m.m.ml}$ when dealing with very small proportion values p than $\hat{p}_{m.m.ml}$, this advantage can be more frequently observed for $\hat{p}_{t.m.ml}$. On the other hand, when the population proportion values approach to 0.5, the efficiency of $\hat{p}_{m.m.ml}$ is considerably higher than all its competitors. This is more noticeable for the set size $m = 5$.

The non-parametric RSS-based estimators are more robust to changes in the proportion values p. In cases where the value of p is close to zero, the performance of

non-parametric estimators are slightly better comparing to the ML proportion estimators. Among the non-parametric estimators, the isotonized estimator $\hat{p}_{m.iso}$ has a slightly higher efficiency. This advantage holds even for proportion values close to zero, and lower correlation level between the concomitants and the binary response variable, where the precision of ML estimators decline dramatically. It is also detectable that the relative efficiency of the two proportion estimators $\hat{p}_{m.st}$, and $\hat{p}_{m.ml}$, that ignore the tie structure among the units, are approximately identical for $\rho = 0.5$, but for better ranking quality the ML estimator $\hat{p}_{m.ml}$ noticeably out performs the standard non-parametric estimator $\hat{p}_{m.st}$. Besides, for good ranking quality, and when fewer number of ties are declared in each set (smaller values of c), the efficiencies of the non-parametric estimators $\hat{p}_{m.iso}$, and $\hat{p}_{m.sp}$ are very close to the top performer $\hat{p}_{m.m.ml}$.

We can clearly notice that over the entire scenarios, the four estimators that utilize the tie structure among the units, achieve substantially higher relative efficiencies by incorporating the ranking data of multiple concomitants. This improvement is more significant when the set size $m = 5$, especially for larger values of ρ. When adding more concomitants into estimation, the increase in efficiency of the mixed ML estimator $\hat{p}_{m.m.ml}$ is appreciable even for correlation coefficient $\rho = 0.5$. It is fascinating to detect that, by combining the ranking information of three concomitants with poor ranking ability, and for proportion values close to 0.5, the estimator $\hat{p}_{m.m.ml}$ has reached the relative efficiency of 2, implying that we can reach the same precision level of $\hat{p}_{SRS}$, but with half of its sample size. Another key observation is that by combining the ranking abilities of concomitants with moderate ranking quality ($\rho = 0.7$), we achieved competitive estimation efficiencies with proportion estimators that only utilize the tie structure information of a single concomitant with a very strong ranking ability ($\rho = 0.9$). This advantage is of vital importance regarding the low frequency of cases where we have access to such powerful concomitant variables.

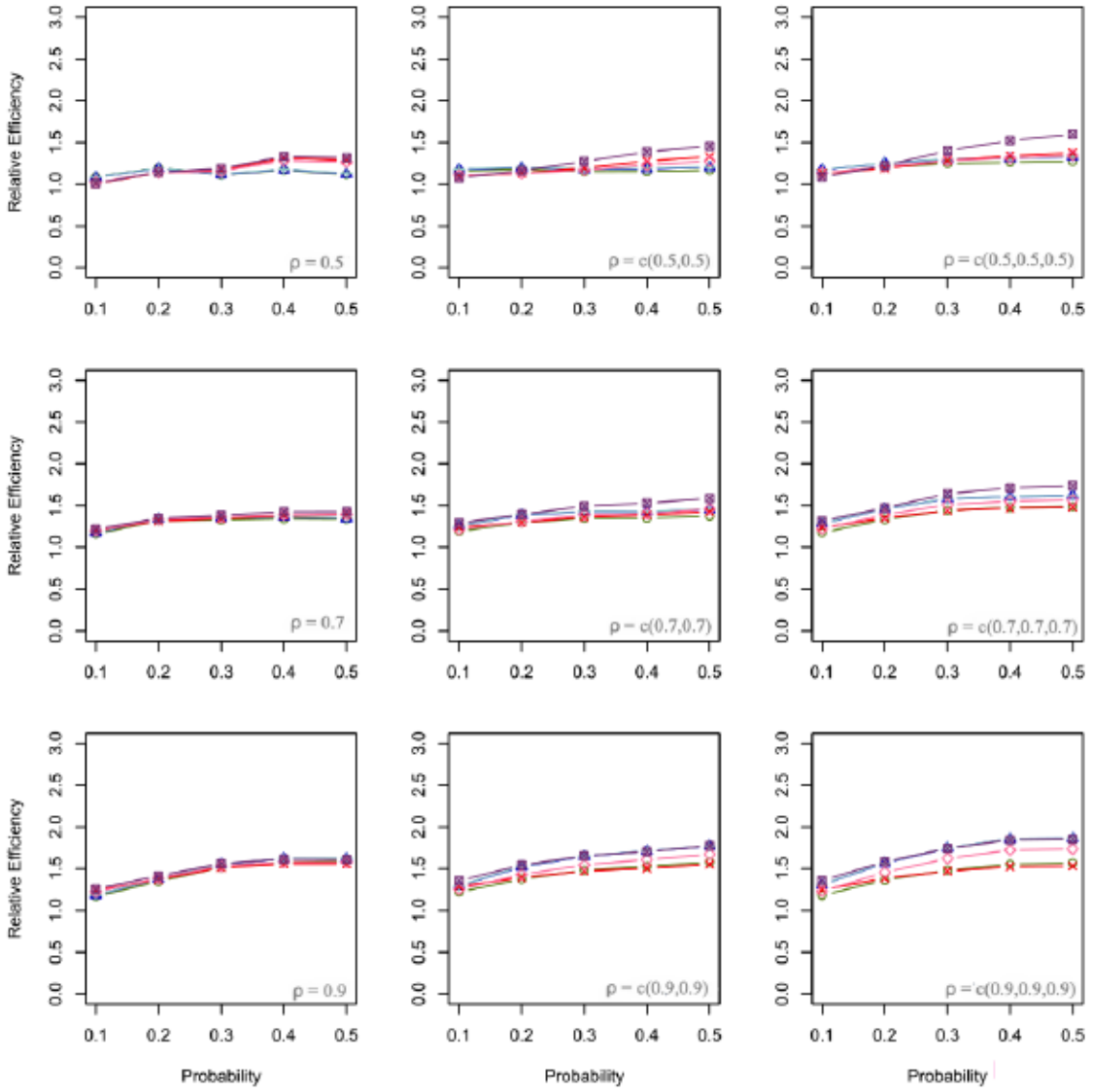

Figure 4.2: Relative efficiencies of $\hat{p}_{m.st}$ (represented by ○), $\hat{p}_{m.sp}$ (represented by △), $\hat{p}_{m.iso}$ (represented by +), $\hat{p}_{m.ml}$ (represented by ×), $\hat{p}_{t.m.ml}$ (represented by ◇), $\hat{p}_{m.m.ml}$ (represented by ⊠) using the TIC model with parameter $c = 0.5$, set size $m = 3$, and $n = 5$ number of cycles.

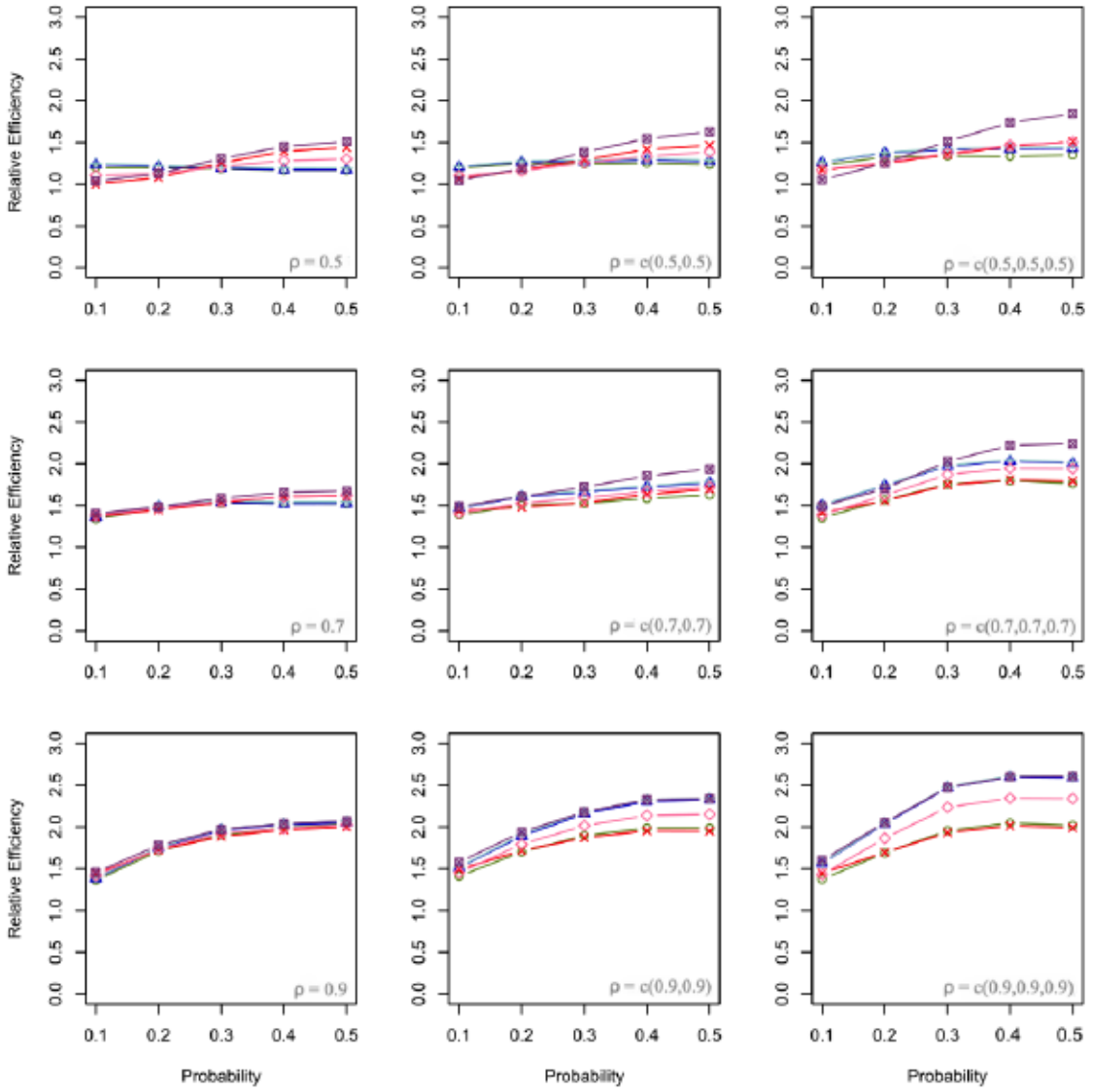

Figure 4.3: Relative efficiencies of $\hat{p}_{m.st}$ (represented by ○), $\hat{p}_{m.sp}$ (represented by △), $\hat{p}_{m.iso}$ (represented by +), $\hat{p}_{m.ml}$ (represented by ×), $\hat{p}_{t.m.ml}$ (represented by ◇), $\hat{p}_{m.m.ml}$ (represented by ⊠) using the DPS model with parameter $c = 0.5$, set size $m = 5$, and $n = 3$ number of cycles.

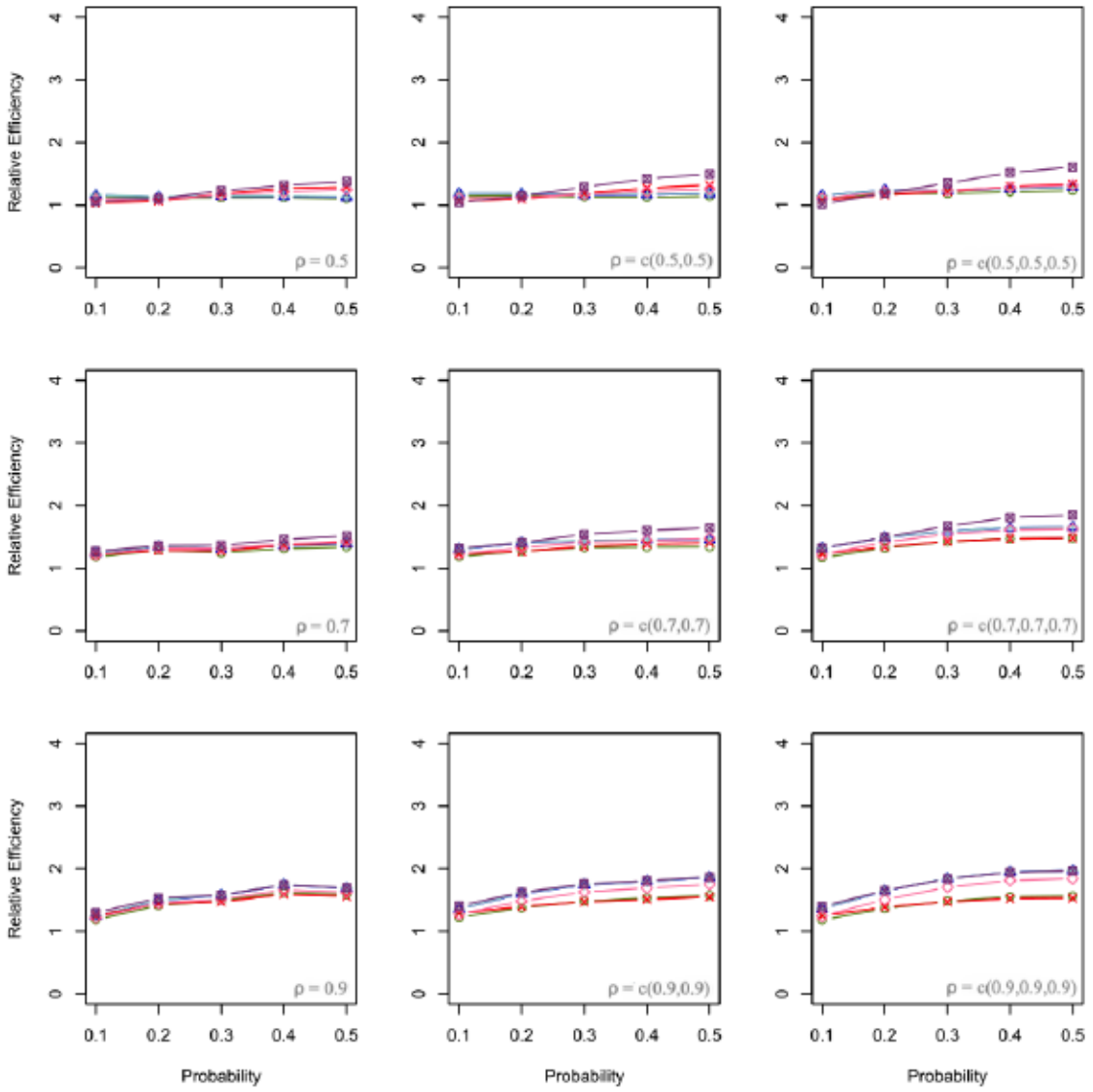

Figure 4.4: Relative efficiencies of $\hat{p}_{m.st}$ (represented by ○), $\hat{p}_{m.sp}$ (represented by △), $\hat{p}_{m.iso}$ (represented by +), $\hat{p}_{m.ml}$ (represented by ×), $\hat{p}_{t.m.ml}$ (represented by ◇), $\hat{p}_{m.m.ml}$ (represented by ⊠) using the TIC model with parameter $c = 0.5$, set size $m = 3$, and $n = 5$ number of cycles.

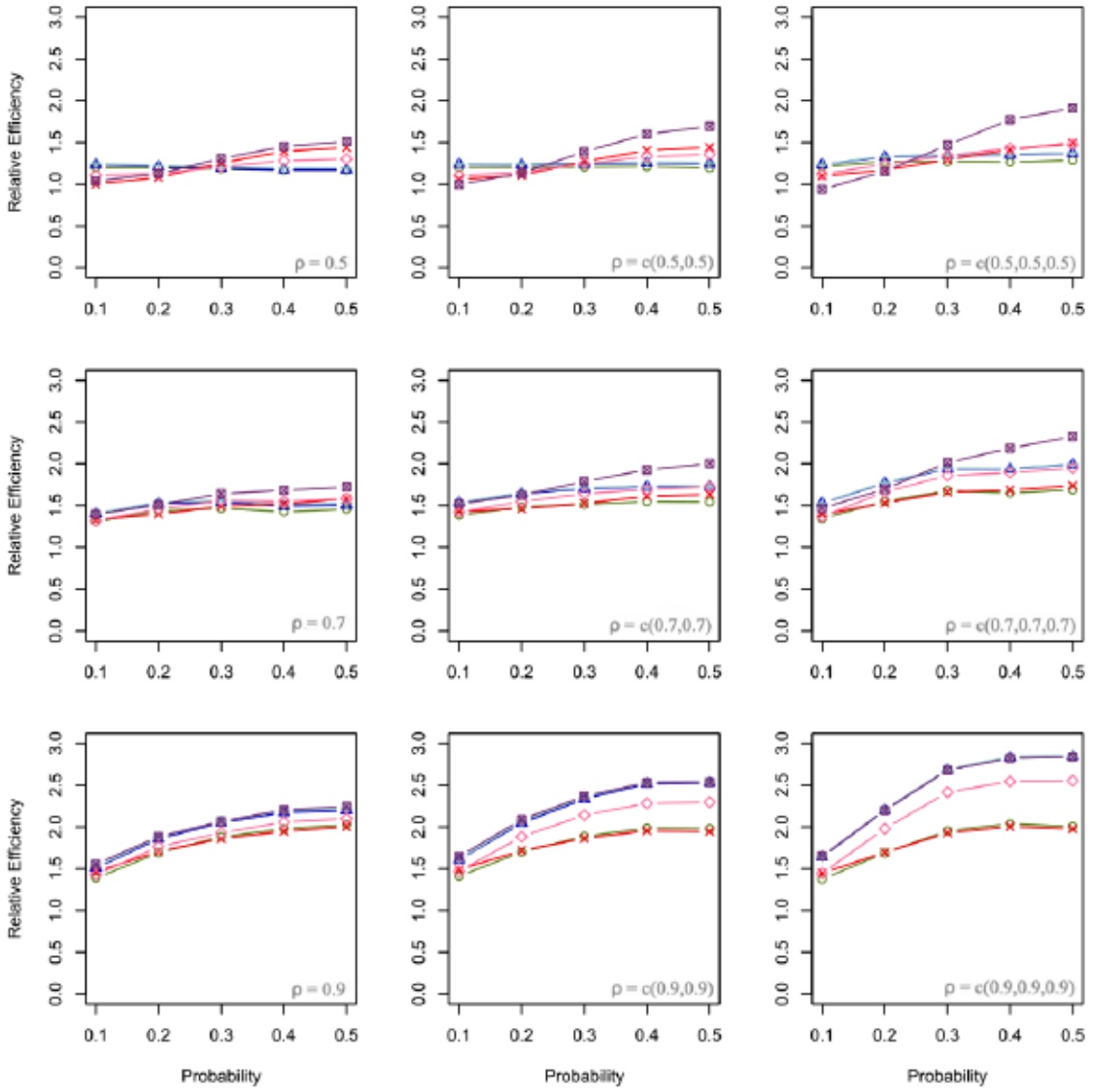

Figure 4.5: Relative efficiencies of $\hat{p}_{m.st}$ (represented by ○), $\hat{p}_{m.sp}$ (represented by △), $\hat{p}_{m.iso}$ (represented by +), $\hat{p}_{m.ml}$ (represented by ×), $\hat{p}_{t.m.ml}$ (represented by ◇), $\hat{p}_{m.m.ml}$ (represented by ⊠) using the TIC model with parameter $c = 0.5$, set size $m = 5$, and $n = 3$ number of cycles.

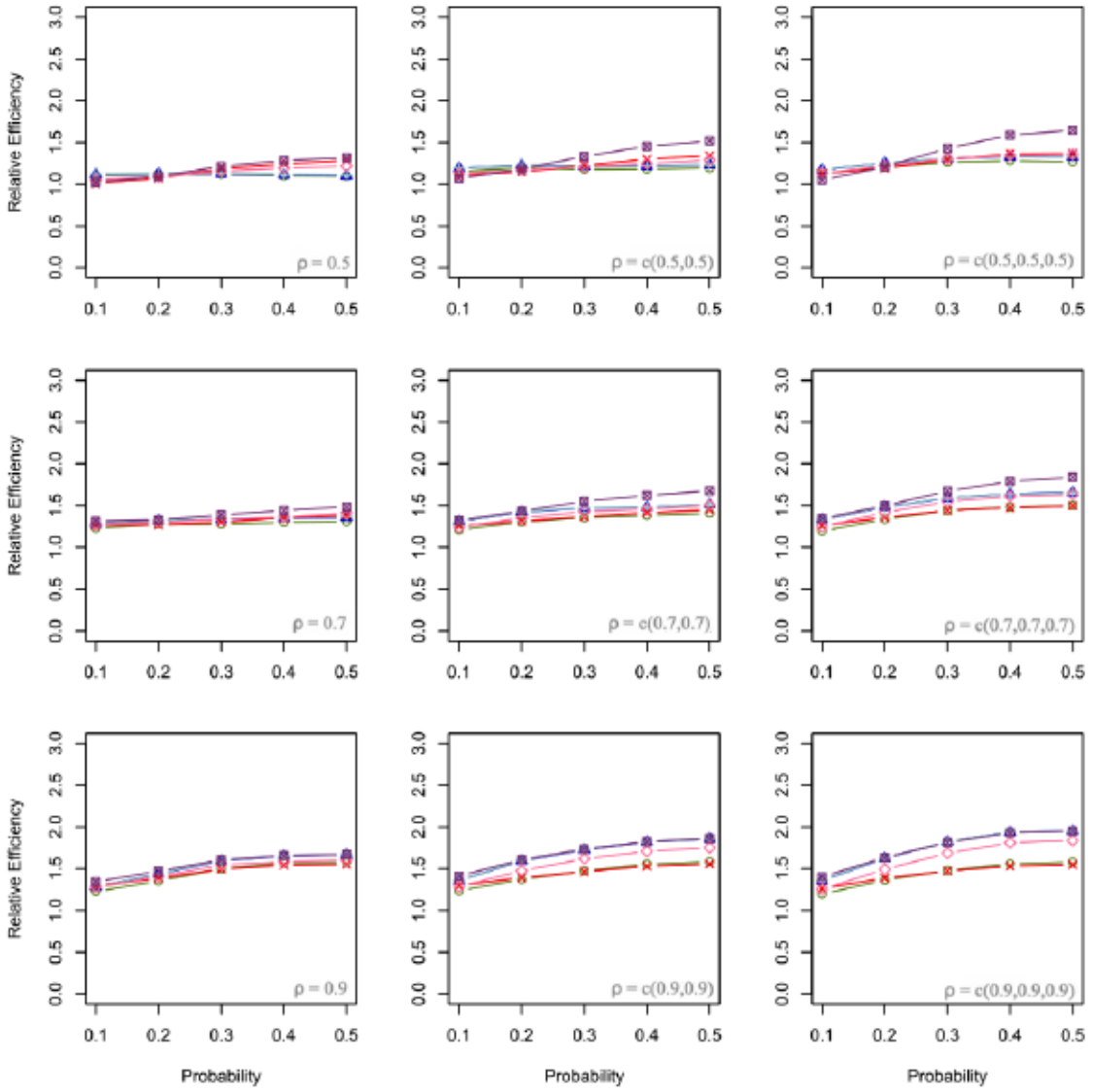

Figure 4.6: Relative efficiencies of $\hat{p}_{m.st}$ (represented by ○), $\hat{p}_{m.sp}$ (represented by △), $\hat{p}_{m.iso}$ (represented by +), $\hat{p}_{m.ml}$ (represented by ×), $\hat{p}_{t.m.ml}$ (represented by ◇), $\hat{p}_{m.m.ml}$ (represented by ⊠) using the DPS model with parameter $c = 1$, set size $m = 3$, and $n = 5$ number of cycles.

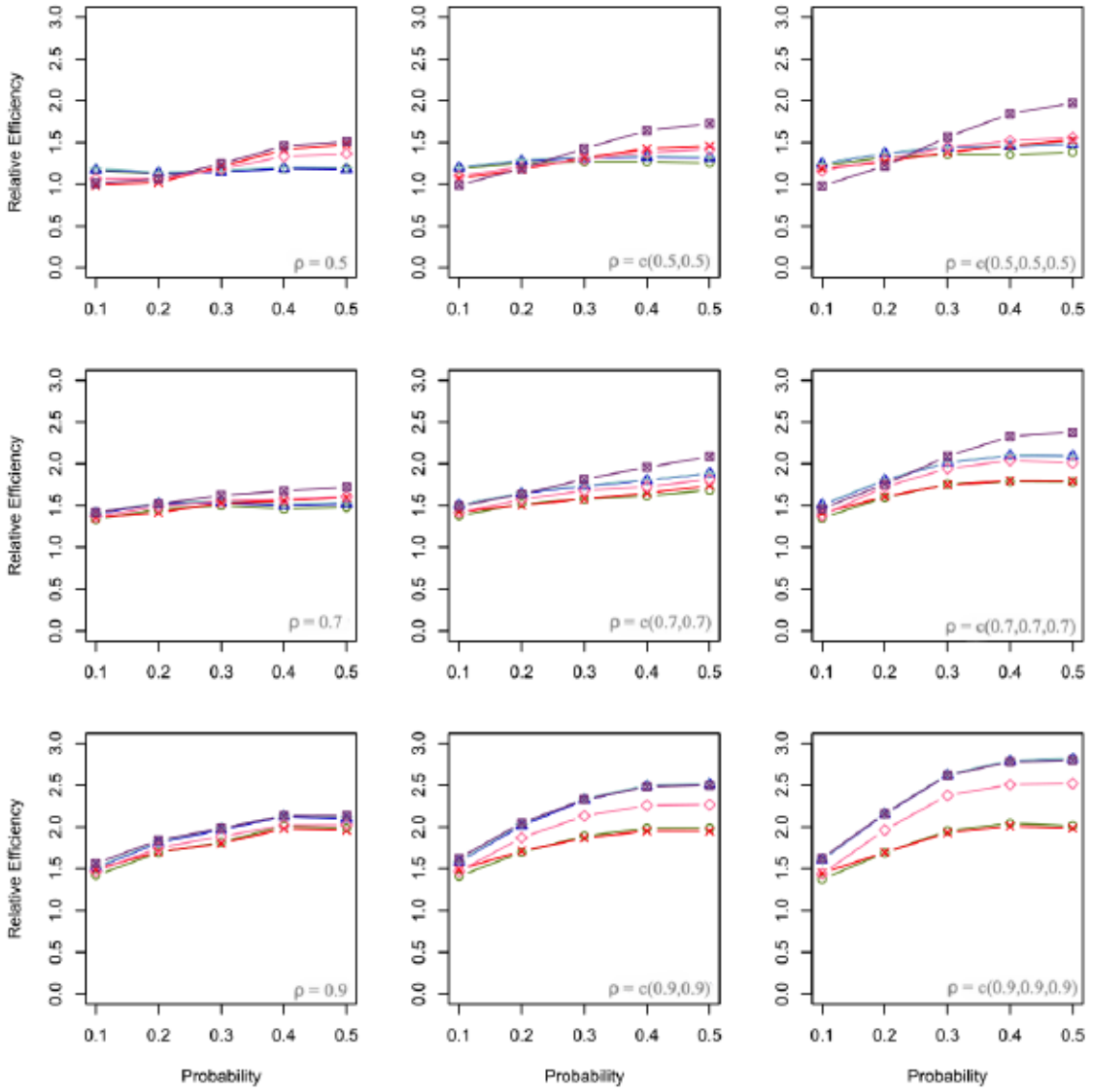

Figure 4.7: Relative efficiencies of $\hat{p}_{m.st}$ (represented by ○), $\hat{p}_{m.sp}$ (represented by △), $\hat{p}_{m.iso}$ (represented by +), $\hat{p}_{m.ml}$ (represented by ×), $\hat{p}_{t.m.ml}$ (represented by ◇), $\hat{p}_{m.m.ml}$ (represented by ⊠) using the DPS model with parameter $c = 1$, set size $m = 5$, and $n = 3$ number of cycles.

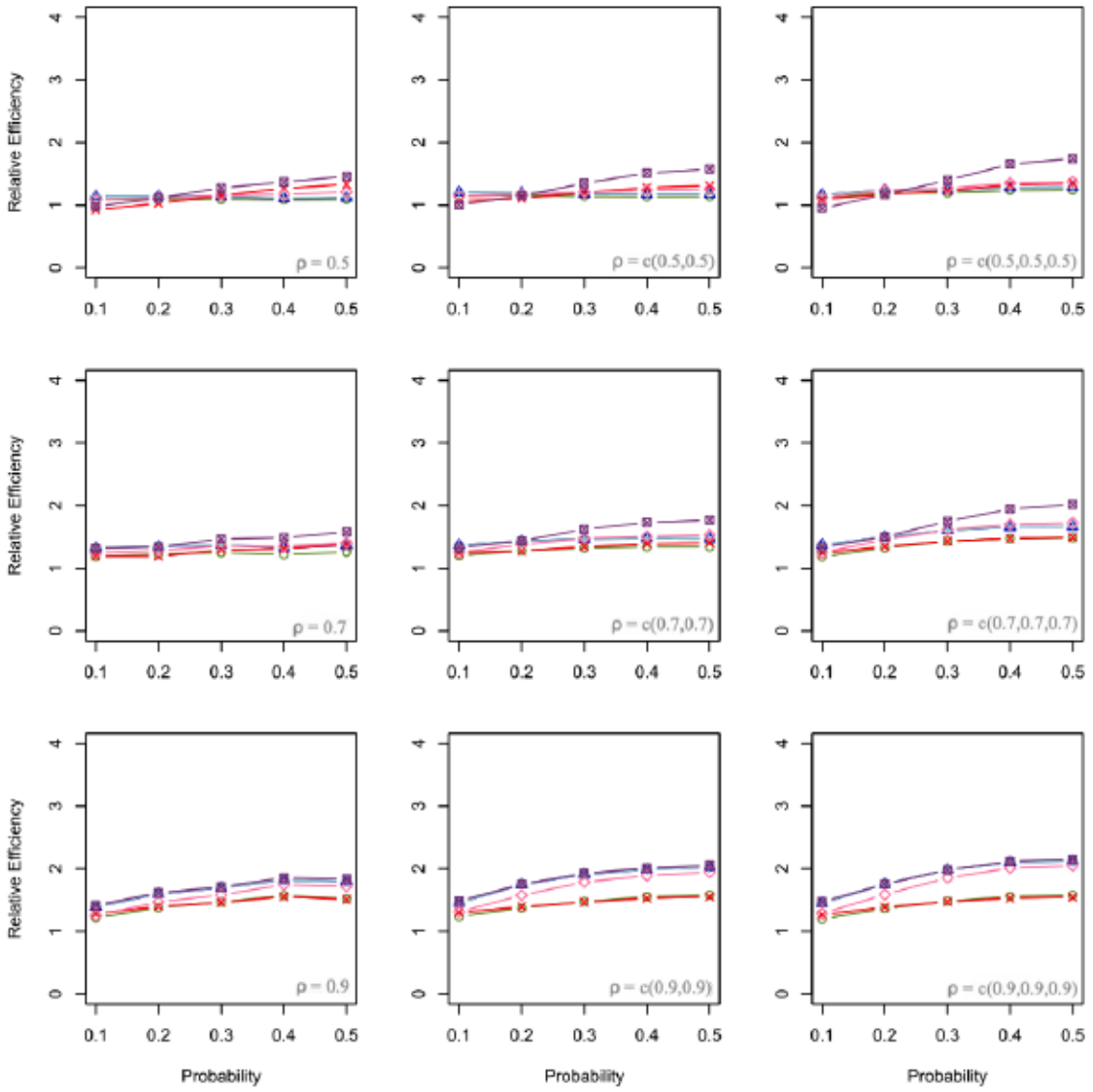

Figure 4.8: Relative efficiencies of $\hat{p}_{m.st}$ (represented by ○), $\hat{p}_{m.sp}$ (represented by △), $\hat{p}_{m.iso}$ (represented by +), $\hat{p}_{m.ml}$ (represented by ×), $\hat{p}_{t.m.ml}$ (represented by ◇), $\hat{p}_{m.m.ml}$ (represented by ⊠) using the TIC model with parameter $c = 1$, set size $m = 3$, and $n = 5$ number of cycles.

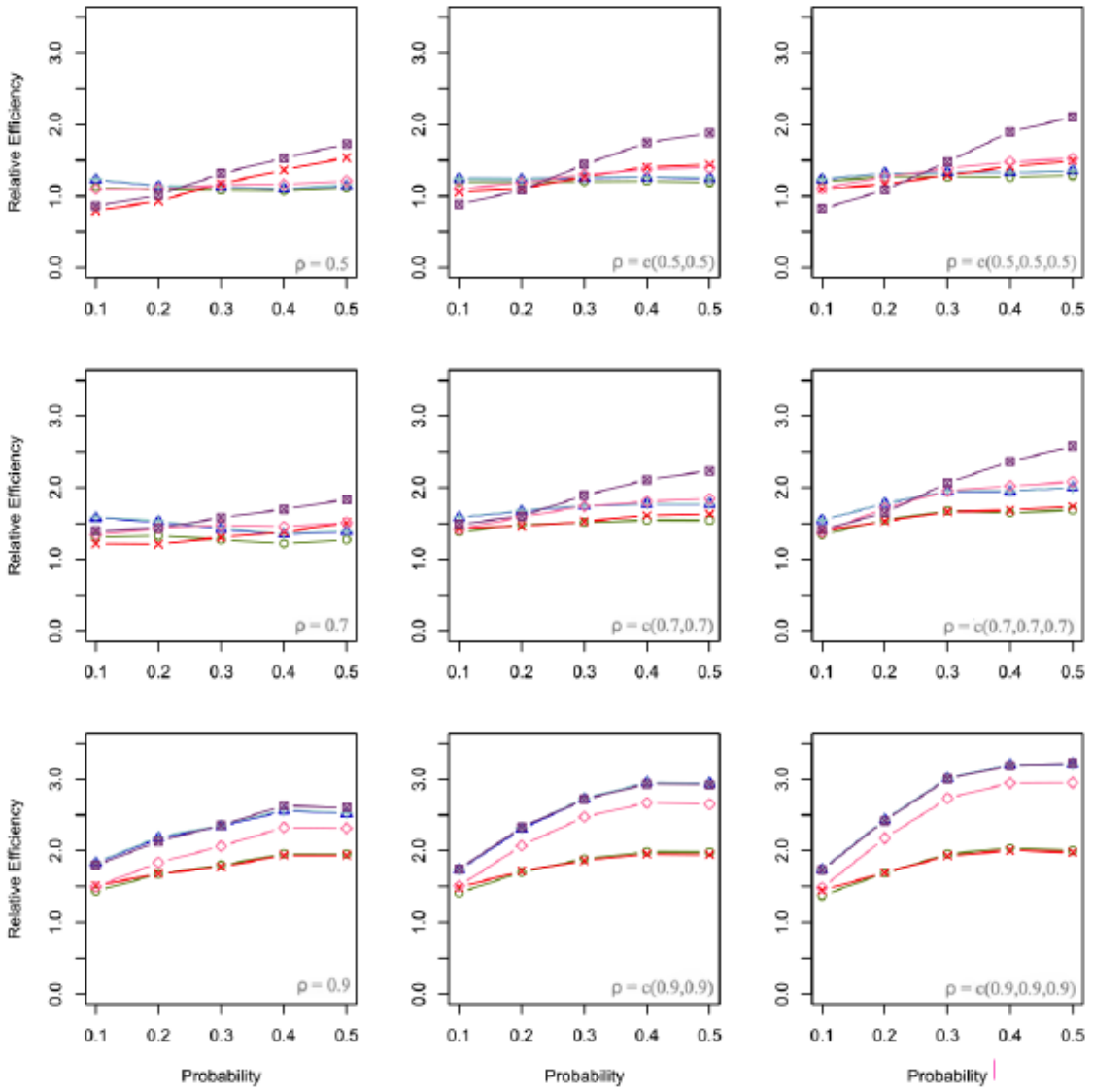

Figure 4.9: Relative efficiencies of $\hat{p}_{m.st}$ (represented by ○), $\hat{p}_{m.sp}$ (represented by △), $\hat{p}_{m.iso}$ (represented by +), $\hat{p}_{m.ml}$ (represented by ×), $\hat{p}_{t.m.ml}$ (represented by ◇), $\hat{p}_{m.m.ml}$ (represented by ⊠) using the TIC model with parameter $c = 1$, set size $m = 5$, and $n = 3$ number of cycles.

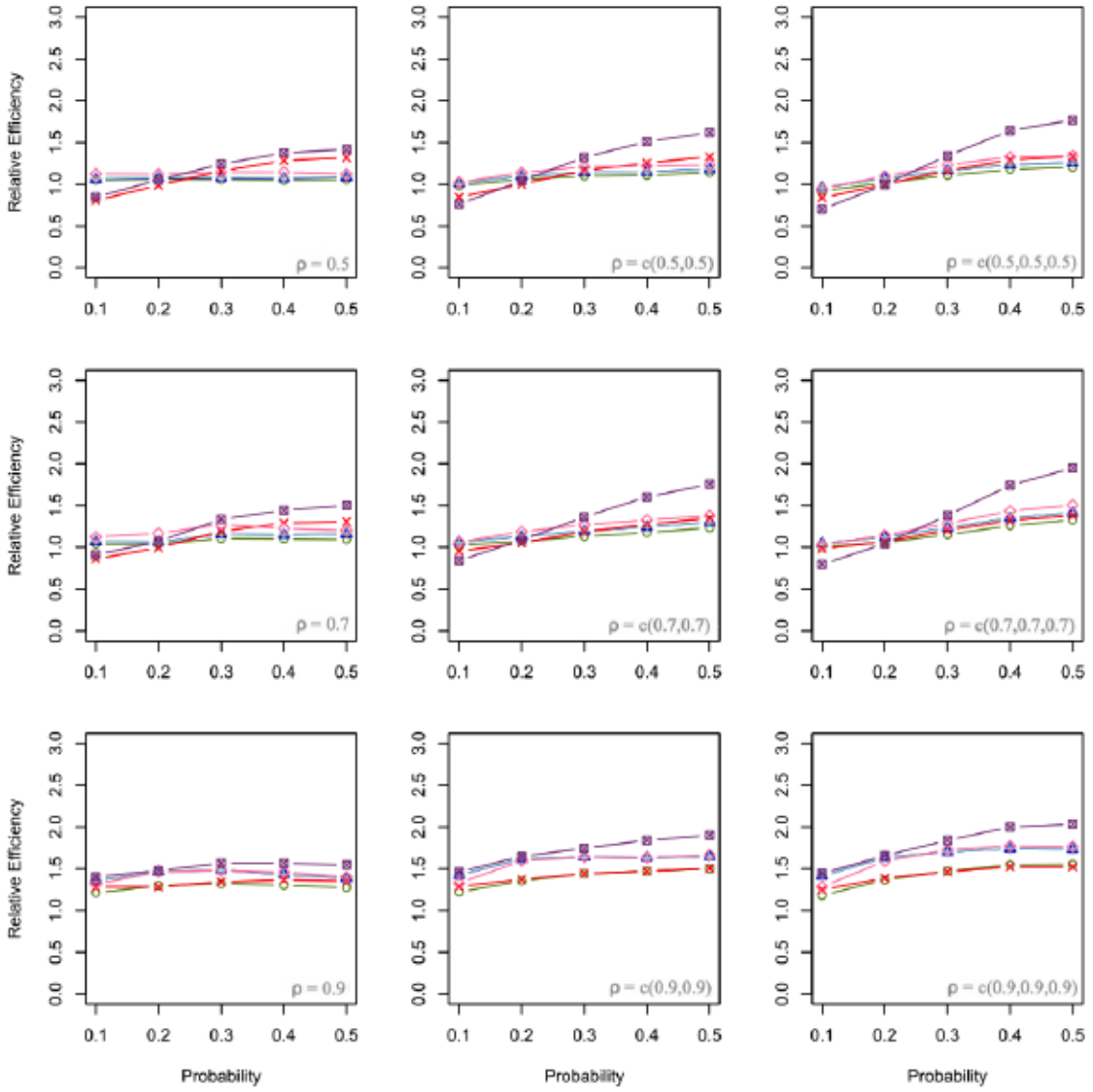

Figure 4.10: Relative efficiencies of $\hat{p}_{m.st}$ (represented by ○), $\hat{p}_{m.sp}$ (represented by △), $\hat{p}_{m.iso}$ (represented by +), $\hat{p}_{m.ml}$ (represented by ×), $\hat{p}_{t.m.ml}$ (represented by ◇), $\hat{p}_{m.m.ml}$ (represented by ⊠) using the DPS model with parameter $c = 4$, set size $m = 3$, and $n = 5$ number of cycles.

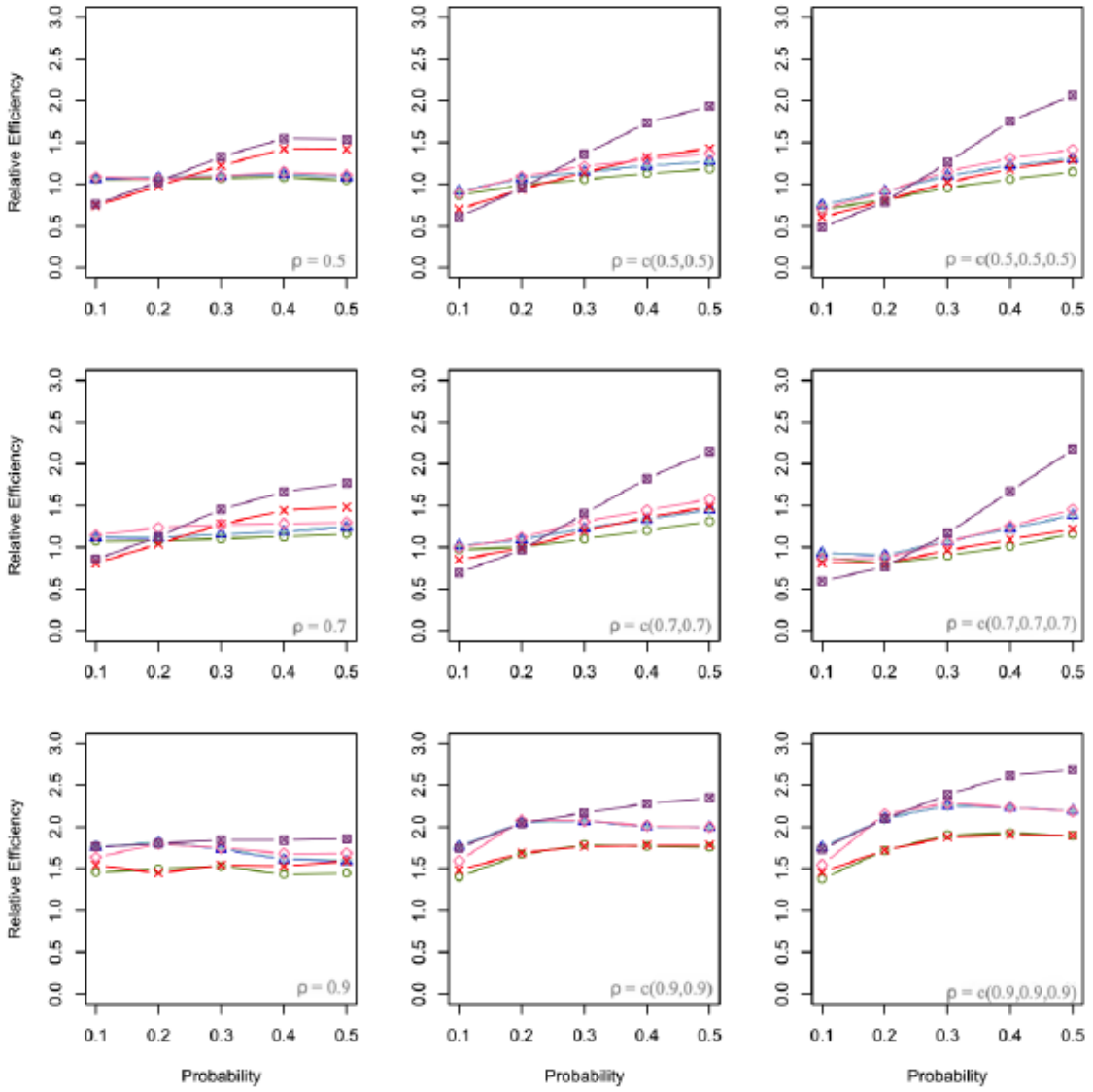

Figure 4.11: Relative efficiencies of $\hat{p}_{m.st}$ (represented by ○), $\hat{p}_{m.sp}$ (represented by △), $\hat{p}_{m.iso}$ (represented by +), $\hat{p}_{m.ml}$ (represented by ×), $\hat{p}_{t.m.ml}$ (represented by ◇), $\hat{p}_{m.m.ml}$ (represented by ⊠) using the DPS model with parameter $c = 4$, set size $m = 5$, and $n = 3$ number of cycles.

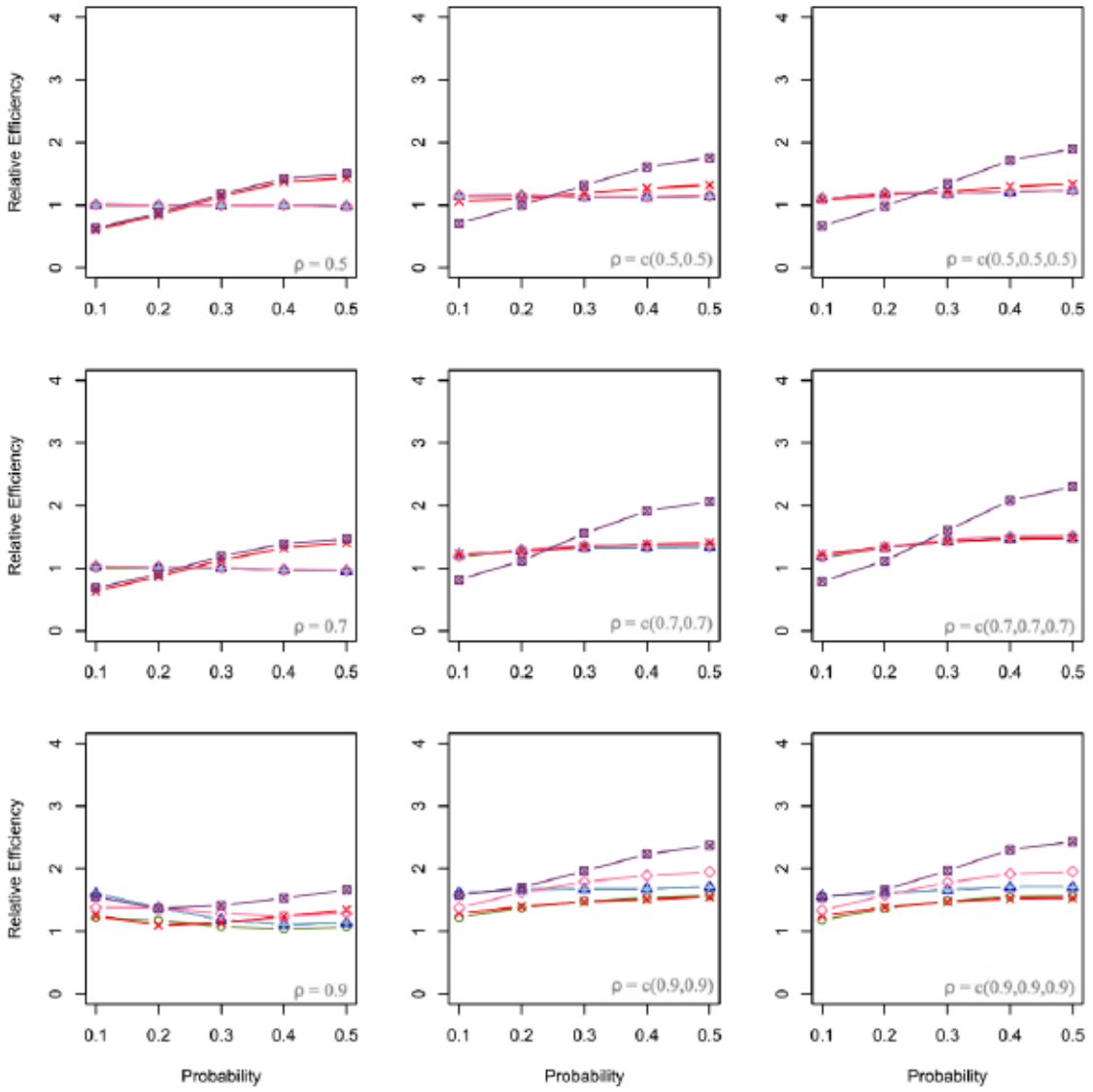

Figure 4.12: Relative efficiencies of $\hat{p}_{m.st}$ (represented by ○), $\hat{p}_{m.sp}$ (represented by △), $\hat{p}_{m.iso}$ (represented by +), $\hat{p}_{m.ml}$ (represented by ×), $\hat{p}_{t.m.ml}$ (represented by ◇), $\hat{p}_{m.m.ml}$ (represented by ⊠) using the TIC model with parameter $c = 4$, set size $m = 3$, and $n = 5$ number of cycles.

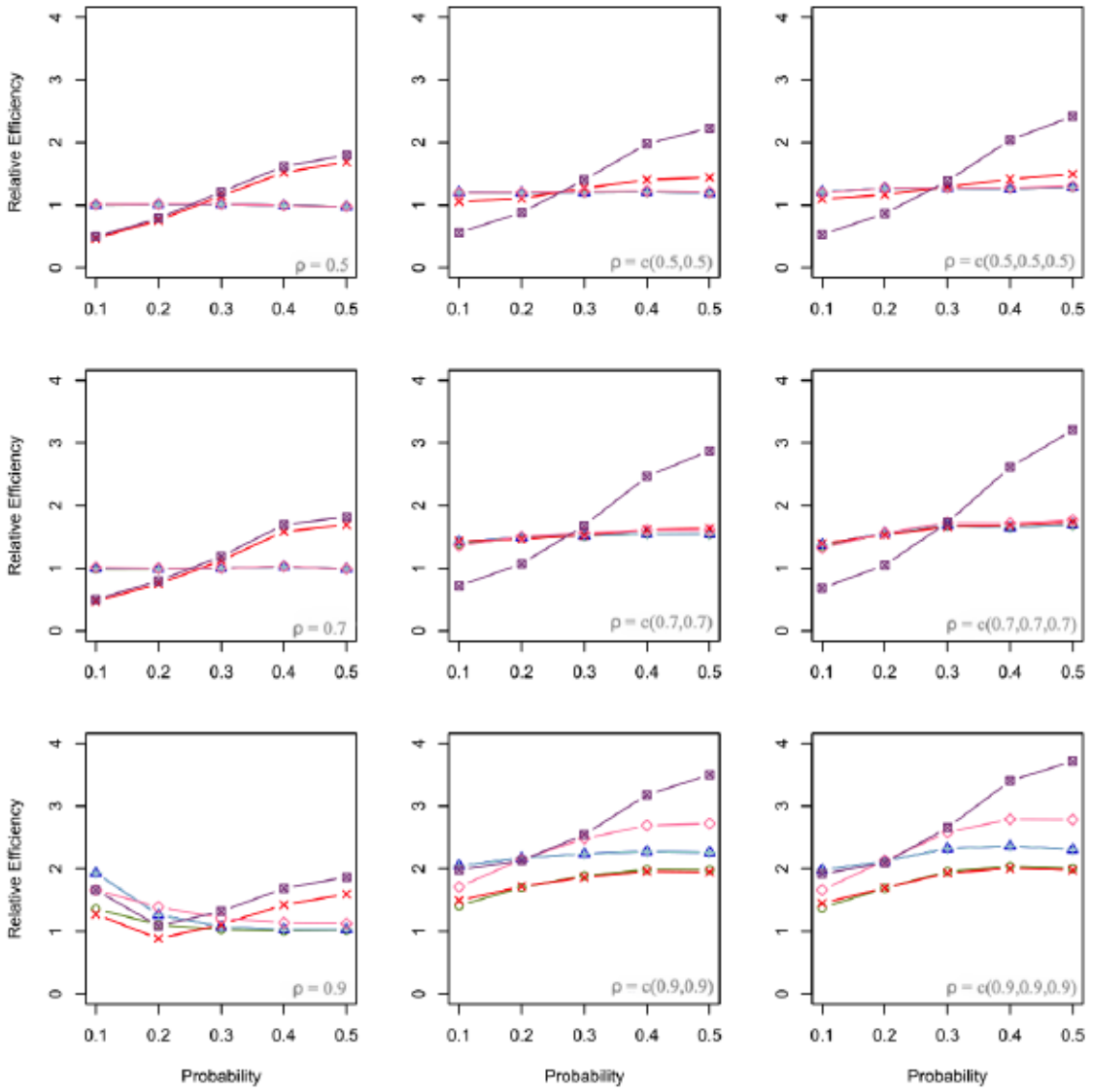

Figure 4.13: Relative efficiencies of $\hat{p}_{m.st}$ (represented by ○), $\hat{p}_{m.sp}$ (represented by △), $\hat{p}_{m.iso}$ (represented by +), $\hat{p}_{m.ml}$ (represented by ×), $\hat{p}_{t.m.ml}$ (represented by ◇), $\hat{p}_{m.m.ml}$ (represented by ⊠) using the TIC model with parameter $c = 4$, set size $m = 5$, and $n = 3$ number of cycles.

4.4 Real Data Analysis

Through the previous sections, by conducting extensive simulation studies, we showed the superiority of various RSS-based non-parametric, and ML estimators of population proportion. According to the cost-effectiveness of the rank-based sampling, we are going to apply the proposed RSS estimators to WBCD data-set to estimate the prevalence of breast cancer in women. As described in Section 1.6, to determine whether a breast tumor is malignant or benign, a comprehensive biopsy procedure is needed. The earlier a person is diagnosed with breast cancer, the higher her chance is to respond positively to treatments, but it may take several months for a person to be detected with breast cancer using this expensive biopsy procedure. Nonetheless, nine visually assessed cytological characteristics associated with breast clumps can be easily obtained through Fine Needle Aspiration (FNA) method. Those concomitant variables are listed in Table 1.2. Now, if we consider malignancy of breast clumps as our binary variable of interest, a set of patients can be ordered using any combination of these cytological characteristics as ranking criterions.

Now, we compare the relative efficiencies of the eight RSS estimators based on the ranking models in Table 4.1, to estimate the breast cancer disease prevalence among the 699 patients in the WBCD data-set, where the true proportion is $p = 0.344$. To measure the impact of different set size m, we again provide the results for $m = \{3, 5\}$ when the total sample size is fixed to be $N = 15$. For each of these cases, we produce 20,000 ranked set samples, and from 4.7, we calculate the efficiencies of these RSS-based estimators with respect to their counterpart based on simple random sampling.

Tables 4.2, and 4.3 demonstrate the relative efficiencies of the eight RSS-based population proportion estimators for the two different set size, and the number of cycles settings. It may be obviously seen that for the entire cases, even by utilizing the

Table 4.1: Ranking models including different combinations of cytological concomitants

Ranking models	Concomitant variables (Correlation level with breast cancer status)
Model 1	Bare nuclei (0.823)
Model 2	Bare nuclei (0.823), Unif. of cell shape (0.823)
Model 3	Bare nuclei (0.823), Unif. of cell shape (0.823), Unif. of cell size (0.821)
Model 4	Normal nucleoli (0.719)
Model 5	Normal nucleoli (0.719), Clump thickness (0.715)
Model 6	Normal nucleoli (0.719), Clump thickness (0.715), Marginal adhe. (0.706)
Model 7	Subject ID (-0.085)

Table 4.2: Relative efficiencies of RSS estimators for different ranking models when the set size $m = 3$, and the number of cycles $n = 5$.

Ranking models	Proportion estimators							
	$\hat{p}_{m.st}$	$\hat{p}_{m.sp}$	$\hat{p}_{m.iso}$	$\hat{p}_{m.ml}$	$\hat{p}_{t.m.ml}$	$\hat{p}_{m.m.ml}$	$\hat{p}_{pros}$	$\hat{p}_{l.reg}$
Model 1	1.396	1.643	1.651	1.409	1.411	1.684	1.352	1.123
Model 2	1.511	1.863	1.868	1.505	1.608	1.885	1.421	1.983
Model 3	1.519	1.871	1.883	1.511	1.634	1.894	1.437	1.192
Model 4	1.273	1.426	1.439	1.324	1.274	1.513	1.232	1.094
Model 5	1.424	1.577	1.576	1.424	1.428	1.650	1.213	1.127
Model 6	1.482	1.673	1.671	1.478	1.568	1.794	1.308	1.170
Model 7	1.021	1.024	1.038	1.334	1.333	1.321	1.049	1.031

subject ID values for ranking the units, the RSS estimators have relative efficiencies greater that one, indicating their superiority over simple random sampling. One may also observe that the mixed ML estimator $\hat{p}_{m.m.ml}$ has a considerably higher estimation precision among its competitors. Furthermore, by adding concomitants with approximately same correlation coefficient with the malignancy of breast tumor, the efficiency of the nearly all proportion estimators improve significantly. However, as we also observed in the numerical studies of Section 4.2, when including more concomitants into the ranking procedure, the efficiency of the proportion estimator $\hat{p}_{l.reg}$ decreases in some cases. This problem is caused by the requirement of training samples by the logistic regression model. It is also evident from the results that, when rankings are performed in the sets of size $m = 5$, the entire of estimators have higher relative efficiencies, however, this should not lead us to conclude a uniformly positive correlation between the value of set size, and the estimation precision. Since by using large sets of units, we are actually making the ranking process more challenging.

Table 4.3: Relative efficiencies of RSS estimators for different ranking models when the set size $m = 5$, and the number of cycles $n = 3$.

Ranking models	Proportion estimators							
	$\hat{p}_{m.st}$	$\hat{p}_{m.sp}$	$\hat{p}_{m.iso}$	$\hat{p}_{m.ml}$	$\hat{p}_{t.m.ml}$	$\hat{p}_{m.m.ml}$	$\hat{p}_{pros}$	$\hat{p}_{l.reg}$
Model 1	1.644	2.017	2.026	1.639	1.619	1.994	1.446	1.182
Model 2	1.869	2.478	2.485	1.845	1.973	2.423	1.617	1.289
Model 3	1.880	2.507	2.509	1.8644	1.994	2.452	1.754	1.303
Model 4	1.425	1.613	1.614	1.418	1.368	1.618	1.295	1.134
Model 5	1.676	1.915	1.927	1.667	1.695	1.934	1.383	1.211
Model 6	1.821	2.203	2.204	1.793	2.018	2.242	1.455	1.292
Model 7	1.024	1.026	1.014	1.403	1.401	1.408	1.029	1.043

Here, we do not provide the bias values since all these RSS estimators are almost unbiased. However, the reader can find the bias values of the $\hat{p}_{l.reg}$, and $\hat{p}_{pros}$ for different set sizes in Hatefi and Jafari Jozani[10], and for the six non-parametric and ML estimators, Zamanzadeh and Wang[28] provided the bias values for single concomitant estimation.

In this chapter, we carried out extensive numerical studies to examine the performance of the previously proposed estimators under various settings of ranking ability, proportion values, tie structures, and set sizes, and also different number of concomitant variables. In addition, we applied the eight RSS-based population proportion estimators to the WBCD data-set introduced in Chapter 1 to estimated the proportion of patients with malignant breast tumors. In both cases, we observed an advantage in the performance of estimators based on RSS data over their counterpart using simple random sampling method. Furthermore, we demonstrated a significant improvement in the precision level of these estimators by incorporating the tie structure data from multiple concomitants.

Chapter 5

Summary and Future Work

5.1 Summary

In studying breast cancer, the tumor status determination (e.g. benign or malignant), requires a comprehensive biopsy procedure which is expensive and time-consuming. Despite this challenge, one has access to visually assessed cytological characteristics (e.g. uniformity of cell shape) that can be obtained easily through fine needle aspiration technique in the Doctors office. In this thesis, we focused on Ranked Set Sampling (RSS) design as one of the most renowned cost-effective sampling techniques. While benefiting from the properties of RSS data, we employed this sampling design in estimation of population proportion. By ranking the individuals from the population based on some concomitant variables (e.g. cytological characteristics), we attempt to estimate the disease prevalence more efficiently.

In Chapter 2, we explored the Partially Rank Ordered Sets (PROS) method through different subsetting techniques introduced by Ozturk [18] as one of the solutions to obtain the tie structure among the individuals for estimation purposes. Accordingly, we discussed a population proportion estimator proposed by Hatefi and Jafari Jozani [10] by focusing on the balanced subsetting through the entire groups of subsets. Although, this balanced PROS proportion estimator can benefit from the tie structure data, the restrictions in the number and size of subsets considerably hinders its performance. Besides, we discussed another proportion estimator studied by Chen et al. [2] that utilizes logistic regression estimates as the ranking criterion. However,

the efficiency of this estimation procedure is impeded owing to the requirement of training samples to estimate the parameters of logistics regression model.

In Chapter 3, upon describing two tie structure classes introduced by Frey [7] that we use in our numerical studies, by providing an illustrative example, we explained how we can use the tie splitting strategy proposed by MacEachern [13] to achieve a combined allocation of observation values. Then, through employing this tie splitting strategy, we proposed an upgraded version of the six non-parametric and Maximum Likelihood (ML) population proportion estimators studied by Zamanzadeh and Wang [28], in such way that they can incorporate the tie structure information obtained from several concomitants. This data then will be used both for selecting sampling units, and also allocating the observed values to the entire of rank strata for which the selected unit has been declared tied.

To investigate the performance of the proposed estimators, we carried out an extensive numerical study under various settings of ranking ability, proportion values, tie structures, and set size values. In both simulation studies, and real data analysis, we used relative efficiencies of the eight RSS based estimators to their counterpart based on simple random sampling as the performance criterion.

Under virtually all the scenarios, we observed relative efficiencies greater that one which implies the advantage of RSS-based estimators relative to their counterpart utilizing SRS data. It was also evident that the value of set size has a significant impact on the estimation precision. This effect is caused by imposing a larger number of rank strata on the underlying population when we divide the sampling units into larger sets, and increasing the chance of obtaining more representative samples. The upgrade in the performance of the PROS estimators is also due to the less amount of restrictions in the construction of subsets for obtaining a PROS data for larger values for the set size. By focusing on the impact of increasing the set size value in logistic regression estimation results, we can see that since this estimator requires training sample with measurements on both the response variable and the concomitant variables for estimating the logistic regression model parameters, the advantage is relatively insignificant.

It is also evident from the results that the performance of all three likelihood-based (ML) estimators are considerably sensitive to changes in the proportion values to the extent that for proportion values close to $p = 0.5$ the ML estimators substantially

outperform their competitors based on RSS data. However, for small proportion values the performance of all ML estimators decline dramatically, and consequently the non-parametric estimators are preferred. Furthermore, among the non-parametric proportion estimators, the isotonized estimator has a slightly higher efficiency over nearly all the cases.

By monitoring the effect of utilizing a combined tie structure information from multiple concomitants, we may detect a noticeable improvement in the performance of estimators that incorporate the tie structure among the sampling units. However, this increase in the estimation efficiency is more significant for larger set sizes, and good ranking quality (larger values of ρ). It is very important to notice that under nearly all the scenarios, by employing a combined tie structure data from multiple concomitants with moderate ranking abilities, we managed to achieve even higher precision level relative to the estimation based on a highly correlated single concomitant. The significance of this achievement is more evident noting that we are often deprived of highly correlated concomitant variable to the costly measurable variable of interest. Also, one may easily observe that by combining the ranking data obtained from the three concomitant variables with the highest correlation level to the disease status of patients, we manage to reach a remarkable increase in the precision level of the proposed estimators.

5.2 Future Work

In this thesis, we only considered population proportion estimators under balanced ranked set sampling. However, in many cases, by an unequal and optimal allocation of sampling units (unbalanced RSS), one can achieve higher estimation precision. Several studies have shown that among a variety of allocation methods, the Neyman allocation will lead to a uniformly more efficient estimators relative to their counterparts based on balanced ranked set sampling (see, e.g., Chen et al. [3]). Through this model, we allocate the number of sampling units to each rank strata proportional to its standard deviation. In the future, we will utilize this allocation model for more efficient estimation of population proportion for different tie structure models, and by employing multiple concomitant variables.

Bibliography

[1] Guvenc Arslan and O Ozturk. "Parametric inference based on partially rank ordered set samples". In: *Journal of the Indian Statistical Association* 51 (2013), pp. 1–24.

[2] Haiying Chen, Elizabeth A Stasny, and Douglas A Wolfe. "Ranked set sampling for efficient estimation of a population proportion". In: *Statistics in medicine* 24.21 (2005), pp. 3319–3329.

[3] Haiying Chen, Elizabeth A Stasny, and Douglas A Wolfe. "Unbalanced ranked set sampling for estimating a population proportion". In: *Biometrics* 62.1 (2006), pp. 150–158.

[4] Min Chen et al. "Generalized isotonized mean estimators for judgment post-stratification with multiple rankers". In: *Journal of agricultural, biological, and environmental statistics* 19.4 (2014), pp. 405–418.

[5] TR Dell and JL Clutter. "Ranked set sampling theory with order statistics background". In: *Biometrics* (1972), pp. 545–555.

[6] Michael A Fligner and Steven N MacEachern. "Nonparametric two-sample methods for ranked-set sample data". In: *Journal of the American Statistical Association* 101.475 (2006), pp. 1107–1118.

[7] Jesse Frey. "Nonparametric mean estimation using partially ordered sets". In: *Environmental and ecological statistics* 19.3 (2012), pp. 309–326.

[8] Jinguo Gao and Omer Ozturk. "Two sample distribution-free inference based on partially rank-ordered set samples". In: *Statistics & Probability Letters* 82.5 (2012), pp. 876–884.

[9] Lowell K Halls and Tommy R Dell. "Trial of ranked-set sampling for forage yields". In: *Forest Science* 12.1 (1966), pp. 22–26.

[10] Armin Hatefi and Mohammad Jafari Jozani. "An improved procedure for estimation of malignant breast cancer prevalence using partially rank ordered set samples with multiple concomitants". In: *Statistical methods in medical research* 26.6 (2017), pp. 2552–2566.

[11] Armin Hatefi, Mohammad Jafari Jozani, and Omer Ozturk. "Mixture Model Analysis of Partially Rank-Ordered Set Samples: Age Groups of Fish from Length-Frequency Data". In: *Scandinavian Journal of Statistics* 42.3 (2015), pp. 848–871.

[12] Johan Lim et al. "Kernel density estimator from ranked set samples". In: *Communications in Statistics-Theory and Methods* 43.10-12 (2014), pp. 2156–2168.

[13] Steven N MacEachern, Elizabeth A Stasny, and Douglas A Wolfe. "Judgement post-stratification with imprecise rankings". In: *Biometrics* 60.1 (2004), pp. 207–215.

[14] GA McIntyre. "A method for unbiased selective sampling, using ranked sets". In: *Australian Journal of Agricultural Research* 3.4 (1952), pp. 385–390.

[15] Xiaosheng Mu. "Log-concavity of a Mixture of Beta Distributions". In: *Statistics & Probability Letters* 99 (2015), pp. 125–130.

[16] Omer Ozturk. "Combining ranking information in judgment post stratified and ranked set sampling designs". In: *Environmental and ecological statistics* 19.1 (2012), pp. 73–93.

[17] Omer Ozturk. "Estimation of population mean and total in a finite population setting using multiple auxiliary variables". In: *Journal of Agricultural, Biological, and Environmental Statistics* 19.2 (2014), pp. 161–184.

[18] Omer Ozturk. "Sampling from partially rank-ordered sets". In: *Environmental and Ecological statistics* 18.4 (2011), pp. 757–779.

[19] François Perron and Bimal K Sinha. "Estimation of variance based on a ranked set sample". In: *Journal of statistical planning and inference* 120.1-2 (2004), pp. 21–28.

[20] Lynne Stokes. "Parametric ranked set sampling". In: *Annals of the Institute of Statistical Mathematics* 47.3 (1995), pp. 465–482.

[21] Jeff T Terpstra and Lindsey A Liudahl. "Concomitant-based rank set sampling proportion estimates". In: *Statistics in medicine* 23.13 (2004), pp. 2061–2070.

[22] Jeff T Terpstra and Zachary A Miller. "Exact inference for a population proportion based on a ranked set sample". In: *Communications in Statistics, Simulation and Computation®* 35.1 (2006), pp. 19–26.

[23] Jeff T Terpstra and Ping Wang. "Confidence intervals for a population proportion based on a ranked set sample". In: *Journal of Statistical Computation and Simulation* 78.4 (2008), pp. 351–366.

[24] You-Gan Wang, Zehua Chen, and Jianbin Liu. "General ranked set sampling with cost considerations". In: *Biometrics* 60.2 (2004), pp. 556–561.

[25] You-Gan Wang, Yimin Ye, and David A Milton. "Efficient designs for sampling and subsampling in fisheries research based on ranked sets". In: *ICES Journal of Marine Science* 66.5 (2009), pp. 928–934.

[26] William H Wolberg and Olvi L Mangasarian. "Multisurface method of pattern separation for medical diagnosis applied to breast cytology." In: *Proceedings of the national academy of sciences* 87.23 (1990), pp. 9193–9196.

[27] Douglas A Wolfe. "Ranked set sampling: its relevance and impact on statistical inference". In: *ISRN Probability and Statistics* 2012 (2012).

[28] Ehsan Zamanzade and Xinlei Wang. "Proportion estimation in ranked set sampling in the presence of tie information". In: *Computational Statistics* (2018), pp. 1–18.